उच्च संकाय प्रशिक्षण केंद्र के अंतर्गत

CENTRE OF ADVANCED FACULTY TRAINING

प्रशिक्षण पुस्तिका-II

Training Manual-II

on

कृषि आँकड़ो के मॉडलिंग एवं पूर्वानुमान के लिए सांख्यिकी एवं मशीन लर्निंग तकनीके

Statistical and Machine Learning Techniques for Modeling and Forecasting Agricultural Data

दिसम्बर 20, 2019 - जनवरी 09, 2020

December 20, 2019 - January 09, 2020

पाठ्यक्रम समन्वयक	:	डॉ मृनमय राय
Course Coordinator	:	Dr. Mrinmoy Ray
पाठ्यक्रम सहसमन्वयक	:	श्री शिवस्वामी जी.पी.
Co-Course Coordinator	:	Dr. Shivaswamy G P
पाठ्यक्रम सहसमन्वयक	:	डॉ हरीश कुमार एच.वी.
Co-Course Coordinator	:	Dr. Harish Kumar H V

पूर्वानुमान एवं कृषि प्रणाली मॉडलिंग प्रभाग
भा.कृ.अ.प. - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू , पूसा नई दिल्ली -110012

Division of Forecasting and Agricultural Systems Modeling
ICAR-Indian Agricultural Statistics Research Institute
Library Avenue, Pusa, New Delhi-110012

2019-2020

प्राक्कथन

भा.कृ.अनु.प.-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान देश में कृषि सांख्यिकी, संगणक अनुप्रयोग और जैवसूचना विज्ञान के विषयों में एक प्रमुख संस्थान है। संस्थान सांख्यिकीय आनुवांशिकी, परीक्षण अभिकल्पना, प्रतिदर्शप पद्धतियाँ, बायोमेट्रिक्स, सांख्यिकीय मॉडलिंग, पूर्वानुमान तकनीक, अर्थमिति, संगणक अनुप्रयोग और जैव-सूचना विज्ञान जैसे विभिन्न क्षेत्रों में अनुसंधान और प्रशिक्षण कार्यक्रम आयोजित करने में व्यस्त हैं। सांख्यिकीय मॉडलिंग कृषि में विविध अनुप्रयोगों के कारण अनुसंधान का एक महत्त्वपूर्ण क्षेत्र है तथा नीति निर्माताओं और कृषि वैज्ञानिकों के लिए उपयोगी है। कृषि आँकड़ों के मॉडलिंग एवं पूर्वानुमान के लिए सांख्यिकी एवं मशीन लर्निंग तकनीक नामक प्रशिक्षण कार्यक्रम सिद्धांत और अनुप्रयोगों का एक मिश्रण है। पाठ्यक्रम के अन्तर्गत विभिन्न विषय शामिल किए गए हैं, किन्तु सीमित नहीं है; फज़ी-रेखीय समाश्रयण, लॉज़िस्टिक समाश्रयण, क्वान्टाइल प्रतिगमन, अरैखिय सांख्यिकी मॉडल, फसल पूर्वानुमान तकनीकें, एरिमा और विरिमा काल श्रृंखला मॉडलिंग, फजीकाल श्रृंखला मॉडलिग, इकोनोमेट्रिक मॉडलिग, कांउट डाटा मॉडलिंग, संरचनात्मक काल श्रृंखला मॉडलिंग, अरेखिय काल श्रृंखला मॉडलिंग, कृत्रिम तंत्रिका नेटवर्क, अनुवांशिक एलगोरिथम, सपोर्ट वेक्टर मशीन, हाईब्रिड काल श्रृंखला मॉडलिंग, कार्ट, स्टोकास्टिक वोलेटिलिटी मॉडल, मार्कोव चेन विश्लेषण, बेसियन काल श्रृंखला मॉडल, रिसेम्पलिंग आधारित प्रतिगमन, रिमोट सेंसिंग, जी.आई.एस तथा कृषि में पूर्वानुमान तकनीक मॉडल का अनुप्रयोग इत्यादि।

इस पाठ्यक्रम के संकाय प्रख्यात सांख्यिकीयविद हैं। जो सांख्यिकीय मॉडलिंग के क्षेत्र में निपुण हैं। इसके अलावा, अतिथि संकाय अपने कार्य क्षेत्र में विषय ज्ञाता होने के कारण प्रसिद्ध शोधकर्ता हैं और वे MNCFC, New Delhi; ICAR-IARI, New Delhi; ICAR-IISS Bhopal; B.C.K.V Mohanpur; ICAR-IIRR, Hyderabad; दिल्ली विश्वविद्यालय और केंद्रीय रेशम बोर्ड जैसे प्रतिष्ठित संगठनों से हैं। सदर्भ पुस्तिका प्रतिभागियों के लिए भविष्य में उपयोगी ज्ञान धरोहर के रूप में सहायक होगी । मैं आशा करता हू इस प्रशिक्षण से प्राप्त अनुभव उन्हें अधिक कुशलता से अनुसंधान करने के लिए सक्षम बनाएगें, इस बहुमूल्य संदर्भ पुस्तिका को समय पर तैयार करने के लिए पाठ्यक्रम समन्वयक, पाठ्यक्रम सह-समन्वयक और आयोजन समिति बधाई के पात्र हैं।

तौकीर अहमद

नई दिल्ली
दिसम्बर 20,2019

(तौकीर अहमद)
निदेशक(का), भा.कृ.अनु.प.-भा.कृ.सां.अनु.सं.

FOREWORD

ICAR-Indian Agricultural Statistics Research Institute is a premier Institute in the disciplines of Agricultural Statistics, Computer Application and Bioinformatics in the country. The Institute has been engaged in conducting research and organizing training programmes in various areas, like Statistical Modelling, Forecasting Techniques, Design of Experiments, Sampling Techniques, Statistical Genetics and Genomics, Computer Applications and Bioinformatics. The present training programme on "Statistical and Machine Learning Techniques for Modeling and Forecasting Agricultural Data" has been planned in such a way that it is a blend of theory and applications. The aim of this training programme is to familiarize the Faculty members/ Scientists/ Researchers at various State Agricultural Universities/ ICAR Institutes with statistical techniques along with machine learning based models for modeling and forecasting of agricultural data in order to draw statistically valid inferences and to help them in upgrading the research, teaching and training skills. Moreover The various topics covered under the course include, but not limited to: Logistic Regression, Quantile Regression, Nonlinear Statistical Models, Crop Forecasting Techniques, ARIMA and Hierarchical time-series modelling, STARMA, VARIMA Time-Series Modelling, Fuzzy Time-Series Modelling, Econometric Modeling, Count Data Modeling, Nonlinear Time-Series Modelling, Artificial Neural Network, Recurrent Neural Network, Genetic Algorithms, Support Vector Machine, Hybrid Time Series Modeling, CART, Stochastic Volatility Models, Bayesian Time Series Modeling, Resampling based Regression, Remote Sensing and GIS etc. along with conventional topics.

The faculty for this course comprises eminent statisticians from ICAR-IASRI, well established in the field of Modeling and Forecasting. Besides, the guest faculties are renowned researchers having sound knowledge in their fields of specialization and also are from reputed organizations like MNCFC, New Delhi; ICAR-IARI, New Delhi; ICAR-IISS Bhopal; B.C.K.V Mohanpur; ICAR-IIRR, Hyderabad; University of Delhi, New Delhi and Central Silk Board. The 'Reference Manual' brought out should serve as a useful wealth of knowledge to the participants for their future use. I am sure that the experience gained from this training will enable them to conduct research more efficiently. I wish to complement Course Coordinators and the Organizing Committee for bringing out this valuable document on time.

(Tauqueer Ahmad)
Director (A), ICAR-IASRI

New Delhi
December 20, 2019

आमुख

भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, कृषि सांख्यिकी, संगणक अनुप्रयोगों और जैव सूचना.विज्ञान के क्षेत्रों में उपक्रम अनुसंधान, षिक्षा और प्रशिक्षण के लिए एक प्रमुख राष्ट्रीय संस्थान माना जाता है। संस्थान राष्ट्रीय कृषि अनुसंधान प्रणाली एवं राष्ट्रीय कृषि सांख्यिकी प्रणाली में योगदान करने एवं इन्हें सुदृढ़ बनाने के लिए सलाहकार एवं परामर्ष सेवाएँ प्रदान करने हेतु विभिन्न महत्वपूर्ण योगदान प्रदान कर रहा है जिसका सीधा प्रभाव राष्ट्रीय नीतियों पर पड़ता है, इस कारण संस्थान को गर्वित स्थान प्राप्त है। सूचना प्रौद्योगिकी के क्षेत्र में प्रगति के साथ, संस्थान वर्तमान जरूरतों और कार्य-पद्धति की चुनौतियों तथा कृषि अनुसंधान की गुणवत्ता के लिए अनुकूल परिस्थितियाँ उपलब्ध करा रहा है। कृषि के क्षेत्र में सांख्यिकीय मॉडलिंग और पूर्वानुमान संस्थान में अनुसंधान के महत्त्वपूर्ण विषयों में से एक है। समस्या के महत्त्व को ध्यान में रखते हुए संस्थान के वैज्ञानिक कृषि के विभिन्न उप कार्यक्षेत्रों में पूर्वानुमान के लिए विभिन्न मॉडलिंग विधियों का अध्ययन करने में जुटे हुए हैं।

भा.कृ.अ.प. और राज्य कृषि विश्वविद्यालयों की क्षमता को म जबूत बनाने के लिए और हमारे अनुसंधान को आवष्यकता के अनुसार एवं विष्वस्तर पर प्रतिस्पर्धी बनाने के लिए यह महत्त्वपूर्ण है कि विभिन्न अनुसंधान गतिविधियों में कार्यरत वैज्ञानिकों को कृषि में पूर्वानुमान के संदर्भ में महत्त्वपूर्ण क्षेत्र सांख्यिकीय मॉडलिंग के आधुनिक विकास से अवगत करवाया जाए। इसी संदर्भ में, कृषि सांख्यिकी एवं संगणक अनुप्रयोग में पूर्वानुमान एवं कृषि प्रणली मॉडलिंग प्रभाग में उच्च संकाय प्रशिक्षण कार्यक्रम के अन्तर्गत भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली, षिक्षा प्रभाग, भारतीय कृषि अनुसंधान परिषद्, नई दिल्ली के संरक्षण में 20 दिसम्बर **2019** से 09 जनवरी, **2020** तक ¨कृषि ऑकड़ों के मॉडलिंग एवं पूर्वानुमान के लिए सांख्यिकी एवं मशीन लर्निंग तकनीक¨ प्रशिक्षण का आयोजन कर रहा है। प्रशिक्षण कार्यक्रम का उद्देश्य कृषि क्षेत्र में पूर्वानुमान के लिए वर्तमान सांख्यिकीय मॉडलिंग विधिया तथा कृषि डेटा के मॉडलिंग और पूर्वानुमान के लिए मशीन लर्निंग आधारित मॉडल को विभिन्न सॉफ्टवेयर (**SAS, R, PYTHON,** तथा **STATA**) के माध्यम से विभिन्न राज्य कृषि विश्वविद्यालयों/भारतीय कृषि अनुसंधान परिषद् के संस्थानों में संकाय सदस्यों/वैज्ञानिकों को अवगत करना है। इससे उन्हें अनुसंधान, शिक्षण और प्रशिक्षण में अपनी क्षमताओं को उन्नत करने में सहायता मिलेगी।

पाठ्यक्रम में मॉडलिंग और कृषि डेटा के पूर्वानुमान के लिए सांख्यिकीय और मशीन सीखने की तकनीक दोनों को शामिल करना शामिल है। पाठ्यक्रम में विभिन्न विषय हैं परन्तु सीमित नहीं हैं: फज़ी रेखीय समाश्रयण, लॉज़िस्टिक समाश्रयण, क्वान्टाइल प्रतिगमन, अरैखिय सांख्यिकी मॉडल, फसल पूर्वानुमान तकनीके, एरिमा और पदानुक्रमित समय-श्रृंखला मॉडलिंग, स्टारमा विरिमा काल श्रृखला मॉडलिंग, फजी काल श्रृंखला मॉडलिग, इकोनोमेट्रिक मॉडलिग, कांउट डाटा मॉडलिंग, संरचनात्मक काल श्रृंखला मॉडलिंग, अरेखिय काल श्रृंखला मॉडलिंग, कृत्रिम तंत्रिका नेटवर्क, आवर्तक तंत्रिका नेटवर्क, आनुवांषिक एलगोरिथम, स्पोर्ट वेक्टर मशीन, हाइब्रिड काल श्रृंखजा मॉडलिंग, कारट, रेन्डम फोरेस्ट तकनीक, स्टोकेस्टिक वोलेटिलिटी मॉडल, मार्कोव चैन विश्लेषण, बेसियन काल श्रृंखला मॉडलए रिसेम्पलिंग आधारित प्रतिगमन, रिमोट सेंसिंग, जी.आई.एस तथा कृषि में पूर्वानुमान तकनीक मॉडल का अनुप्रयोग इत्यादि साथ पारंपरिक विषयो के बारे में जानकारी देना है।

हम संस्थान के संकाय और अतिथि संकाय का धन्यवाद करते हैं जिन्होंने अपना बहुमूल्य समय समर्पित कर इस पाठ्यक्रम को सार्थक और सफल बनाने में सहायता की तथा जिनके अथक प्रयासों से यह संदर्भ.पुस्तिका समय पर तैयार हो सकी। इस प्रशिक्षण का आयोजन करने के लिए षिक्षा प्रभाग भारतीय कृषि अनुसंधान परिषद्, नई दिल्ली द्वारा प्रदान की गई आवष्यक धन राषि के लिए हम धन्यवाद करते हैं। हम विभिन्न भा.कृ.अ.प. संस्थानों और राज्य कृषि विश्वविद्यालयों के आभारी हैं जिन्होंने अपने वैज्ञानिको/प्रोफेसरों को इस प्रशिक्षण में प्रतिभागिता हेतु नियुक्त किया। हम डॉ. तौकीर अहमद, निदेशक, भा.कृ.सां.अ.सं. के कृतज्ञ हैं, जिन्होंने हमें इस पाठ्यक्रम को आयोजित करने का उत्तरदायित्व सौंपा। हम डॉ. के.एन. सिंह, प्रधान पूर्वानुमान एवं कृषि प्रणाली मॉडलिंग प्रभाग के आभारी हैं जिन्होनें बहुमूल्य मार्गदर्षन किया और पाठ्यक्रम के सुचारू रूप से संचालन के लिए आवष्यक सुविधाएँ उपलब्ध करवाइंर्। इस प्रशिक्षण कार्यक्रम से जुड़े विभिन्न पहलुओं और गतिविधियों को सरल बनाने में सहायता करने के लिए प्रषासनिक, वित्तीय, तकनीकी और र ञायक स्टाफ के कर्मचारियों के लिए हम कृतज्ञता प्रकट करते हैं। अंत में, हम उन सभी का धन्यवाद करते हैं जिन्होनें प्रत्यक्ष या परोक्ष रूप में इस संदर्भ.पुस्तिका को तैयार करने में सहायता प्रदान की।

नई दिल्ली
दिसम्बर 20, 2019

(मृनमय राय)
पाठ्यक्रम समन्वयक

(शिवस्वामी जी.पी.)
पाठ्यक्रम सहसमन्वयक

(हरीश कुमार एच.वी.)
पाठ्यक्रम सहसमन्वयक

PREFACE

The institute occupies a place of pride in providing advisory and consultancy services to support and strengthen National Agricultural Research System as well as National Agricultural Statistics System by making several significant contributions which have a direct impact on the national policies. With the advances in information technology, the institute has all along been adapting itself to the current needs and methodological challenges and quality enrichment of agricultural research. Statistical modeling and Forecasting in the domain of agriculture is one of the important subjects of research at the Institute. Considering the importance of the problem, scientists in the institute are engaged in studying various modeling approaches for their forecasting applications in different sub domains of agriculture.

As a capacity strengthening initiative in ICAR institutes and SAUs and in order to make our research need based and globally competitive, it is important that the scientists engaged in various research activities are exposed to the latest developments taking place in the important area of statistical modeling in the context of forecasting in agriculture. Forecasting and Agricultural Systems Modeling division of ICAR-IASRI is organizing Centre of Advanced Faculty Training (CAFT) programme "Statistical and Machine Learning Techniques for Modeling and Forecasting Agricultural Data" from 20th December, 2019 to 09th January, 2020 at ICAR-IASRI, New Delhi under the aegis of Education Division, ICAR, New Delhi. The aim of the training programme is to provide exposure to Faculty members/ Scientists at various State Agricultural Universities/ ICAR institutes on current Statistical modeling methodologies along with machine learning based models for modeling and forecasting of agricultural data through use of various software packages (SAS, R, PYTHON and STATA) with particular emphasis on applications in agriculture. This would help them in upgrading their capabilities in research, teaching and training.

The course is structured to include both statistical and machine learning techniques for modelling and forecasting of agricultural data. The various topics covered under the course include, but not limited to: Logistic Regression, Quantile Regression, Nonlinear Statistical Models, Crop Forecasting Techniques, ARIMA and Hierarchical time-series modelling, STARMA, VARIMA Time-Series Modelling, Fuzzy Time-Series Modelling, Econometric Modeling, Count data Modeling, Nonlinear Time-Series Modelling, Artificial Neural Network, Recurrent Neural Network, Genetic algorithms, Support Vector Machine, Hybrid Time Series Modeling, CART, Stochastic volatility models, Bayesian Time Series Modeling, Remote Sensing and GIS, Applications of Technology Forecasting models in agriculture etc. along with conventional topics.

We take this opportunity to thank the faculty of the institute and the guest faculty who devoted their valuable time in making this course meaningful and successful and whose efforts helped in bringing out this manual on time. Necessary funds provided by Education Division, ICAR, New Delhi for conducting this training are duly acknowledged. We are also thankful to the various ICAR Institutes and State Agricultural Universities for deputing their scientists/ professors to this course. We are indebted to Dr. Tauqueer Ahmad, Director (Acting), ICAR-IASRI for entrusting us with the responsibility of organizing this course. We are also thankful to Dr. K N Singh, Head, Division of Forecasting and Agricultural Systems Modeling for his valuable guidance and making all necessary facilities available for smooth conduct of the course. We also place on record our thankfulness to the F&ASM division staff, administrative, financial, technical and auxiliary staff for their wholehearted support in facilitating various items and activities for this training. Finally, we are thankful to one and all, especially who helped us in preparing this manual.

New Delhi **Dec 20, 2019**	**(Mrinmoy Ray)** **Course Coordinator**	**(Shivaswamy G P)** **Course Co-Coordinator**	**(HarishKumar H V)** **Course Co-Coordinator**

CONTENTS

Application of GIS and Remote Sensing for Crop Yield Forecasting

K. N. Singh and Bishal Gurung
ICAR-IASRI, New Delhi
knsingh@icar.gov.in

The soil fertility changes occur due to cropping, manure and fertilizer applications. Soil test results of one farm need to have scope to be connected with the broader population of all farms in a given area. But we may not be able to sample each farm in the population, because it is too costly, troublesome and time consuming, especially with the multiple small farm holdings as in India. We thus need to generalize results over an entire area**.** For the periods between 1975 to 1980, soil fertility maps for nitrogen (N), phosphorous (P) and potassium (K) were prepared using soil test data generated by soil testing laboratories functioned throughout the country (Ghosh and Hasan, 1979). Till date there is no major up-gradation in these maps. Singh et al. (2004) used point estimates for districts to prepare soil fertility maps of N, P and K for the states of Andhra Pradesh and Maharashtra. Further, Singh et al. (2006) have interlinked fertilizer recommendations for targeted yields of crops with these maps. Soil fertility maps have been prepared for 12 agriculturally important states using Soil Index Values (IV) for each district. Index values were calculated using standard procedure (Biswas and Mukherjee, 1987).

The IVs were classified in to three categories viz. (Low 0- 1.5, Medium 1.5-2.5 and High >2.5). Soil Test Crop Response (STCR) approach was used to prescribe optimum doses of nutrients, based on available soil nutrients. From available nutrient Index Values and STCR equations the backward calculation for soil test values (STV) were obtained as follows:

Low :	0.0 - 1.5	::	0-a (a>0)
Medium:	1.5 - 2.5	::	a-b (b>a)
High :	>2.5 ::	>b	

If IV<=1.5
STV= ax (IV)/(1.5)
If IV>1.5 and<= 2.5
STV= a+[(b-a) x (IV-1.5)]
If IV >2.5

STV=(b/2.5) x IV

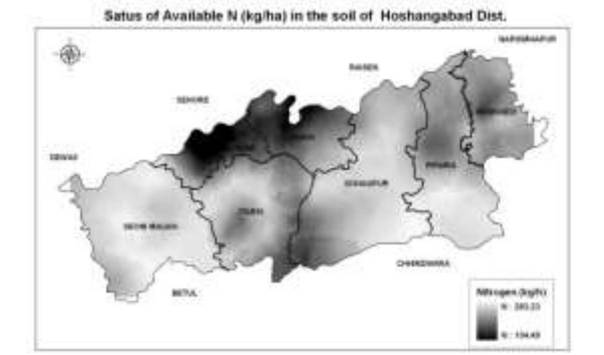
Where a and b were positive coefficients used for describing the range of different nutrients. The a and b values depend on soil characteristics and are different for different soils. These denote the fertility of a soil with respect to N, P or K and are determined through soil test crop response correlation experiments. If a soil sample has available nutrient (N, P or K) below ‘a’ that means it is low, between ‘a’ and ‘b’, it is medium and above ‘b’, it is high. The district wise index values have been assigned from the database generated on N, P and K index values to the corresponding district layer of the state in GIS and generated the thematic maps accordingly.

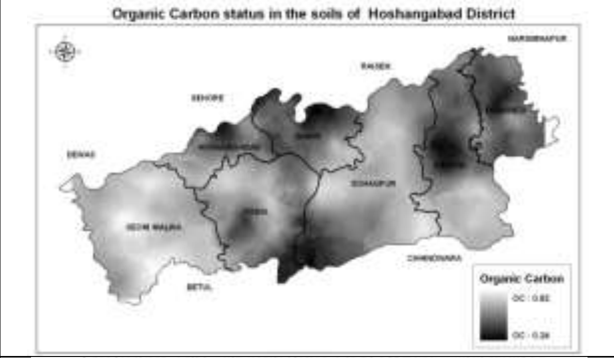
The calculated soil test values were incorporated into the fertility maps to prescribe nutrients for targeted yields. This online application Software was developed to recommend fertilizer doses for the targeted yield at the District level. This system has the facility to input actual soil test values at the farmer’s fields to obtain optimum dozes. The application is a user-friendly tool to help the farmer in improving the efficiency (appropriate dose) of fertilizer use to achieve a specific crop yield.

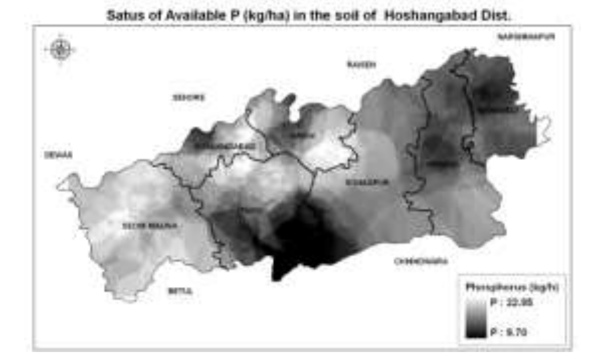
Remote Sensing, Geographic Information Systems (GIS), and the Agricultural Non-point Source Pollution (AGNPS) model have been used to assess runoff and sediment yield from various sub-watersheds above Cheney Reservoir in Kansas, USA (Bhuyan et al. (2002)). Ray and Dadhwal (2001) used satellite-based remote sensing data and GIS tools for estimating seasonal crop evapotranspiration in Mahi Right Bank Canal (MRBC) command area of Gujarat, India.

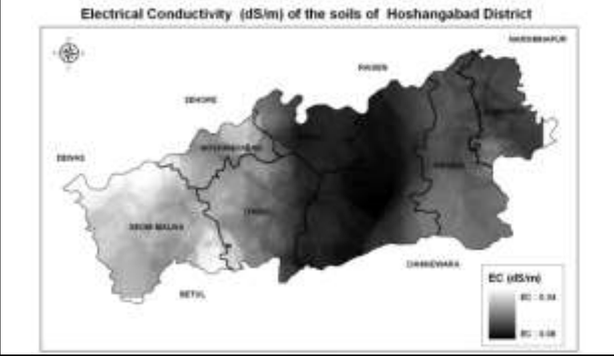
The recent technologies like GIS and GPS thus have much to offer for preparing soil fertility maps. Once the soil fertility maps are created, it is possible to transform the information from Soil Test Crop Response models into Spatial fertilizer recommendation maps. Such maps provide site-specific recommendation, validation for soil fertility over the following years. The fertilizer doses for targeted yield can be prescribed to the farmers by locating his field/ area on the map with the help of latitude/longitude information.

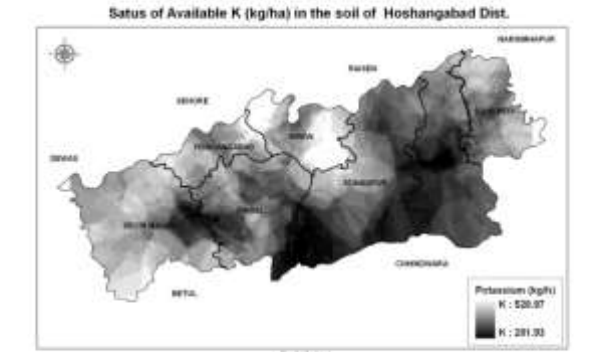
To cover complete district Stratified Multistage Stratified Random Sampling has been adapted**.** To select the soil samples from different categories (big, small and marginal) of farmers it was essential to select farmers and to select farmers first to select village which is first stage unit. To select villages from a tehsil Simple Random Sampling without Replacement (SRSWOR) has been used. There is problem of spatial estimation, sometimes called spatial prediction. This arises in case a spatial field is partially observed at selected sites and the goal is to infer the field at unobserved sites. An example of spatial random field is soil nutrient concentrations over an agricultural domain. Among different methods of spatial interpolation of soil properties, kriging is an optimal interpolation method (Issak and Srivastava, 1989). To select the best model Akaike's (1973) information criterion (AIC) has been used.











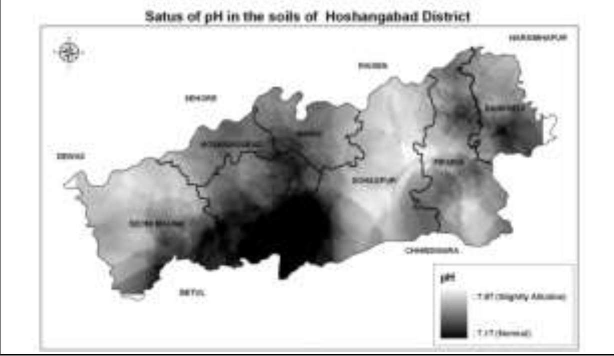


Fig 1. Kriged raster images (response surface) of different soil nutrients of Hoshangabad district.

In case of N, spherical method had the least AIC value. Hence for N, spherical Variogram method was used for kriging. Similarly linear, spherical, exponential, linear and linear Variogram methods of kriging were used for P, K, OC, EC and pH respectively.

Estimated response surface (Fig. 1) clearly showed that in Hoshangabad district OC in soil ranged between 0.28% to 0.81%, available soil N was in the range of 104 to 279 kg/ha, available soil P was in the range of 10 to 22.9 kg/ha, available soil K was in the range of 282 to 529 kg/ha. The EC was in the range of 0.08 to 0.34 desi siemens (dS/m) and pH was in the range of 7.2 to 7.9. With the help of these raster images all the ground points (pixel) was assigned with unique estimated value of respective nutrients. It was observed that calculated Abs (t) was less than that of tabulated t (for $P< 0.05$) for all the nutrients in 2007. This showed that in subsequent year there was no significant change in these nutrients. The results of year 2008 showed that only pH changed. For other nutrients there was no significant difference. Therefore, it is inferred that observed soil parameters for Hoshangabad district did not change significantly for at least two consecutive years except for pH. Again a web based on line spatial fertilizer recommendation system has been developed where farmers can get information up to field level if he has knowledge of Longitude and Latitude, otherwise all the villages have been included in the system and village wise recommendation can be obtained.

The crop yield can be forecasted using the equations of the form:

FN=3.92T–0.46SN, 4.26T-0.59SN, 3.47T-0.37SN, 4.00T-0.44SN, 3.78T- 0.48SN, etc.

FP_2O_5=2.61T- 2.45SP, 2.35T-3.16SP, 2.53T- 2.12SP, 2.32T-2.09SP, 2.39T-2.90SP, etc.

FK_2O =2.47T-0.25SK, 1.89T-0.20SK, 2.12T-0.20SK, 1.82T-0.17SK, 1.24T-0.12SK, etc.

Where,
FN, FP_2O_5, FK_2O are fertilizer applied.

SN, SP, SK are soil test values for N, P and K and T is forecast yield (qt/ha)

Singh et al. (2006) utilized remote sensing data for preparing land productivity maps using simple linear relationship between Normalized Difference Vegetative Index (VDVI) values and Land Productivity Index (LPI) values. Satellite data for selected areas of

Hoshangabad and Guna Districts have been used to obtain relationship between NDVI values and soil nutrients (Singh et al. (2009)). To obtain frequency data for each nutrient polygon to carry out statistical analysis the union of polygons was performed. Relationships between nutrients and NDVI values have been obtained for different months. The results indicate satisfactory relationship between nutrients and NDVI values. There is a good agreement between available nitrogen and maximum of maximum of NDVI values and it is best in the month of December and February. Available phosphorus can be estimated using average of average of NDVI values in the month of February and Potassium can be estimated using average of average of NDVI in the month of December.

Bibliography

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaksi, editors), Akademiai Kiado, Budapest, Hungary, Pp. 267-281

Biswas T. D. and Mukherjee S. K. (1987). Text Book of Soil Science. Tata Mc Graw-Hill Publishing Company Limited, New Delhi. P.193.

Bhuyan, SJ; Marzen, LJ; Koelliker, JK; Harrington, JA Jr.; Barnes, PL (2002). Assessment of runoff and sediment yield using remote sensing, GIS, and AGNPS. Journal of Soil and Water Conservation Ankeny. 57(6), 351-364.

Ghosh, A. B. and Hasan, R. (1979). Bulletin of Indian Society of Soil Science, 12, 1-8.

Issak, E. H. and Srivastava, R. M. (1989). An introduction to Applied Geostatistics, Oxford Univ. Press, New York, p.561

K. N. Singh. N. S. Raju and A. Subba Rao (2006). Land Productivity Assessment using Remote Sensing (RS) and Geographic Information System (GIS). Indian Journal of Agricultural Sciences, 76 (2) 81-84.

K. N. Singh, N. S. Raju, A. Subba Rao, Abhishek Rathore, Sanjay Srivastava, R. K.

Samanta and A. K. Maji (2006). Prescribing optimum doses of nutrients for targeted yield through soil fertility maps in Andhra Pradesh (AP). Jour. Ind. Soc. Agril. Stat. 59(2): 131-140.

K. N. Singh, Abhishek Rathore, A. K. Tripathi, A. Subba Rao, Salman Khan and Bharat Singh (2009). Use of geographic information system, remote sensing and global positioning system in the application of precise fertilizer to maintain soil productivity of the farmers fields. New Technology for rural development having potential of commercialization. Allied Publishers Pvt. Ltd., New Delhi. PP 183-195

Ray, S. S., Dadhwal, V. K. (2001). Estimation of crop evapotranspiration of irrigation command area using remote sensing and GIS. Agricultural-Water-Management. 49(3), 239-249.

Singh, K. N., Raju, N. S., Subba Rao A., Srivastava Sanjay and Maji A. K. (2004). GIS based system for prescribing optimum dose of nutrients for targeted yield through soil fertility maps in Andhra Pradesh (AP). In the Proceedings of National workshop on recent trends in earth resources mapping, MANIT Bhopal. pp. 67-72.

Singh, K. N., Raju, N. S., Subba Rao A., Srivastava Sanjay and Maji A. K. (2004). GIS based system for prescribing optimum dose of nutrients for targeted yield through soil fertility maps in Maharashtra. In the Proceedings of National Seminar on Information and Communication Technology for Agriculture and Rural Development NAARM, Hyderabad. pp. 167-174.

Regression Analysis: Diagnostics and Remedial Measures

Lalmohan Bhar
ICAR-IASRI, New Delhi
lmbhar@iasri.res.in; lmbhar@gmail.com

1. Introduction

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others. This methodology is widely used in business, the social and behavioral sciences, the biological sciences including agriculture. For example, fish weight at harvest can be predicted by utilizing the relationship between fish weights and other growth affecting factors like water temperature, dissolved oxygen, free carbon dioxide etc. There are other situations in agriculture where relationship among variables can be exploited through regression analysis. We frequently use equations to summarize or describe a set of data. Regression analysis is helpful in developing such equations. For example we may collect a considerable amount of fish growth data and data on a number of biotic and abiotic factors, and a regression model would probably be a much more convenient and useful summary of those data than a table or even a graph. Besides prediction, regression models may be used for control purposes.

A functional relation between two variables is expressed by a mathematical formula. If *x* denotes the independent variable and *y* the dependent variable, then *y* can be related *x* through a functional relation of the form *y* = *f*(*x*). Given a particular value of *x*, the function *f* indicates the corresponding value of *y*. In regression analysis, the variable x is known as input variable, explanatory variable or predictor variable. This is an exact mathematical relationship. In statistical relation, may not be perfect owing to sampling. The above functional form is made a statistical model by adding an error term as

$$y = f(x) + \varepsilon ,$$

where *e* denotes the error term.

Depending on the nature of the relationships between *x* and *y*, regression approach may be classified into two broad categories *viz*., linear regression models and nonlinear regression models. The response variable is generally related to other causal variables through some parameters. The models that are linear in these parameters are known as linear models; whereas in nonlinear models parameters appear nonlinearly.

2. Linear Regression Models

We consider a basic linear model where there is only one predictor variable and the regression function is linear. The model can be stated as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1)$$

Where y_i is the value of the response variable in the i^{th} trial β_0 and β_1 are parameters, x_i is the value of the predictor variable in the i^{th} trial, ε_i is a random error term with mean zero and variance σ^2 and ε_i and ε_j are uncorrelated so that their covariance is zero.

Regression model (1) is said to be simple and linear regression model. It is “simple” in the sense that there is only one predictor variable and “linear” in the sense that all parameters appeared linearly with the predictor variables. The parameters β_0 and β_1 in regression model (1) are called regression coefficients, β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of y per unit increase in x. The parameter β_0 is the y intercept of the regression line. When the scope of the model includes $x = 0$, β_0 gives the mean of the probability distribution of y at $x = 0$. When the scope of the model does not cover $x = 0$, β_0 does not have any particular meaning as a separate term in the regression model. Extension of this model to more than one predictor variable is straight forward. If the linear model contains more than one predictor variable, then it is known as multiple linear regression model. For example, if we have p predictor variables, then a multiple linear regression model can be formulated as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i \, . \qquad (2)$$

If the model contains an intercept term, then this model is known as model “with intercept” otherwise it is known as “no intercept” model. Thus both the models (1) and (2) are “with intercept” models. In practice, we must be careful in choosing the model. If our situation demands that there should not be any intercept in the model, then we should use a “no intercept” model. On the other hand if in our situation even after putting the values of predictor variables as zero, we get some response, we should use a “with intercept” model.

Estimation of Parameters

In the above models the variables *y* and *x* are known, these are observed. The only unknown quantities are the parameters*β's.* In regression analysis, our main concern is how precisely we can estimate these parameters. Once these parameters are estimated, our model becomes known and we can use it for further analysis. The method of least squares is generally used to estimate these parameters. For each observations (x_i, y_i), the method of least squares considers the error of each observation, *i.e*, for a simple model $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. The method of least squares requires the sum of the *n* squared errors. This criterion is denoted by *Q*:

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \ . \tag{3}$$

According to the method of least squares, the estimators of β_0 and β_1 are those values $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, that minimize the criterion *Q* for the given observations. To minimize *Q*, we differentiate *Q* with respect to each parameter and equate to zero. We get as many equations as the number of parameters. Solving these equations simultaneously, we get the estimates of parameters. For example for the regression model (1) the values of $\hat{\beta}_0$ and $\hat{\beta}_1$that minimizes *Q* for any particular set of sample data are given by the following simultaneous equations:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \ ,$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \ .$$

These two equations are called normal equations and can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$ as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \ ,$$

$$\beta_0 = \frac{1}{n} (\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i) = \bar{y} - \hat{\beta}_1 \bar{x} \ ,$$

where $\bar{x}$ and $\bar{y}$ are the means of the x_i and the y_i observations, respectively.

Some Properties of Fitted Regression Line

Once the parameters estimates are obtained, the fitted line would be

$$\hat{y}_i = \beta_0 + \beta_1 x_i \tag{4}$$

We can compute residuals of each observation. The *i*th residual is the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$, *i.e.*, $r_i = y_i - \hat{y}_i$. These residuals play an important role in diagnosing any problem associated with data. The estimated regression line (4) fitted by the method of least squares has a number of properties worth noting.

(i) The sum of the residuals is zero, $\sum_{i=1}^{n} r_i = 0$.

(ii) Sum of the squared residuals, $\sum_{i=1}^{n} r_i^2$ is a minimum.

(iii)Sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$.

(iv)Sum of the weighted residuals is zero, weighted by the level of the predictor variable in the *i*th observation, *i.e.*, $\sum_{i=1}^{n} x_i r_i = 0$.

(v) Sum of the weighted residuals is zero, weighted by the fitted value of the response variable in the *i*th observation, *i.e.*, $\sum_{i=1}^{n} \hat{y}_i r_i = 0$.

(vi)The regression line always goes through the points ($\bar{x}$, $\bar{y}$).

Estimation of Error Term Variance σ^2

The variance σ^2 of the error terms ε_i in regression model needs to be estimated to know the variability of the probability distribution of *y*. In addition, a variety of inferences concerning the regression function and the prediction of *y* require an estimate of σ^2. Denote by $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} r_i^2$, is the residual sum of squares. Then an estimate of σ^2 is given by,

$$\hat{\sigma}^2 = \frac{SSE}{n-p}, \tag{5}$$

where p is the total number of parameters involved in the model including the intercept term, if the model contains it. We also denote this quantity by *MSE*.

Inferences in Linear Models

In multiple linear regression model, all variables may not be contributing significantly to the model. In other word, each of the parameters may not be significant. Therefore, these parameters must be tested whether they are significantly different from zero or not. That is, we test the null hypothesis (H_0) against the alternative hypothesis (H_1) for a parameter β_i (say) as follows:

$$H_0 = \beta_i = 0$$

$$H_1 = \beta_i \neq 0.$$

When $H_0 = \beta_i = 0$ is accepted we infer that there is no linear association between y and x_i. For normal error regression model, the condition $\beta_i = 0$ implies even more than no linear association between y and x_i. $\beta_i = 0$ for the normal error regression model implies not only that there is no linear association between y and x_i but also that there is no relation of any kind between y and x_i, since the probability distribution of y are then identical at all levels of x_i. The test is based on t test

$$t = \frac{\beta_i}{s(\beta_i)},$$

where $s(\beta_i)$ is the standard error of β_i and calculated as $s(\beta_i) = \sqrt{\dfrac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$.

The decision rule with this test statistic when controlling level of significance at α is

if $|t| \leq t(1-\alpha/2; n-p)$, conclude H_0,

if $|t| > t(1-\alpha/2; n-p)$, conclude H_1.

Similarly testing for other parameters can be carried out.

Prediction of New Observations

The new observation on y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of x for the new observation as x_h and the new observation on y as y_h. We also assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

The distinction between estimation of the mean response, and prediction of a new response, is basic. In the former case, we estimate the mean of the distribution of y. In the present case, we predict an individual outcome drawn from the distribution of y. The great majority of individual outcomes deviate from the mean response, and this must be taken into account by the procedure for predicting $y_{h(new)}$. We denote by $\hat{y}_h$, the predicted new observation and by $\sigma^2(\hat{y}_h)$ the variance of $\hat{y}_h$. An unbiased estimator of $\sigma^2(\hat{y}_h)$ is given by $\hat{\sigma}^2(\hat{y}_h) = \hat{\sigma}^2 + s^2(\hat{y}_h)$, where $s^2(\hat{y}_h)$ is the estimate of variance of prediction at x_h and given by

$$s^2(\hat{y}_h) = \hat{\sigma}^2(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} . \qquad (6)$$

Confidence interval of $\hat{y}_h$ can be constructed by using t-statistic namely,

$$\hat{y}_h \pm t(1-\alpha/2; n-p)\ \sigma^2(\hat{y}_h) .$$

Measure of Fitting, R^2

The overall fitting of a regression line can be judged by the F-statistic by carrying out an analysis of variance. If the F-statistic is significant, we say that our model is fitted well. However, there are times when the degree of linear association is of interest. A frequently used statistic is R^2. We describe this descriptive measure to describe the degree of linear association between y and x.

Denote by $SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2$, total sum of squares which measures the variation in the observation y_i , or the uncertainty in predicting y, when no account of the predictor variable x is taken. Thus $SSTO$ is a measure of uncertainty in predicting y when x is not considered. Similarly, SSE measures the variation in the y_i when a regression model utilizing the predictor variable x is employed. A natural measure of the effect of x in reducing the variation in y, *i.e.*, in reducing the uncertaintinty in predicting y, is to express the reduction in variation ($SSTO-SSE=SSR$) as a proportion of the total variation and it is denoted by R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad (7)$$

The measure R^2 is called coefficient of determination, $0 \le R^2 \le 1$. In practice R^2 is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between *x* and *y*. Remember that R^2 statistic should be used only when in the model an intercept term is involved. For the model with no intercept, R^2 is not a good statistic. In case of "no intercept" model, sum of all residuals may not be equal to 0, making R^2 inflated.

An Example

Consider the following data:

Table 1: Data

Case No.	x_1	x_2	x_3	y		***Case No.***	x_1	x_2	x_3	y
1	12.980	0.317	9.998	57.702		14	14.231	10.401	1.041	41.896
2	14.295	2.028	6.776	59.296		15	15.222	1.220	6.149	63.264
3	15.531	5.305	2.947	56.166		16	15.740	10.612	-1.691	45.798
4	15.133	4.738	4.201	55.767		17	14.958	4.815	4.111	58.699
5	15.342	7.038	2.053	51.722		18	14.125	3.153	8.453	50.086
6	17.149	5.982	-0.055	60.446		19	16.391	9.698	-1.714	48.890
7	15.462	2.737	4.657	60.715		20	16.452	3.912	2.145	62.213
8	12.801	10.663	3.048	37.447		21	13.535	7.625	3.851	45.625
9	17.039	5.132	0.257	60.974		22	14.199	4.474	5.112	53.923
10	13.172	2.039	8.738	55.270		23	15.837	5.753	2.087	55.799
11	16.125	2.271	2.101	59.289		24	16.565	8.546	8.974	56.741
12	14.340	4.077	5.545	54.027		25	13.322	8.589	4.011	43.145
13	12.923	2.643	9.331	53.199		26	15.949	8.290	-0.248	50.706

In the present example, we have 3 three predictor variables x_1, x_2 and x_3 and there are 26 observations. The response variable denoted by *y*. Applying least square method we obtain the parameter estimates as follows:

Table 2: ANOVA with intercept model

Source	**Degrees of freedom**	**Sum of Square**	**Mean Square**	**F-value**	**Prob. > F**
Model	3	1062.34	354.11	109.69	<.0001
Error	22	71.02	3.22		
Corrected Total	25	1133.37			

Table 3: Parameter Estimates with intercept model

Variable	Degrees of freedom	Parameter Estimates	Standard Error	t-value	Prob. > |t|
Intercept	1	8.19	6.29	1.30	0.2060
x_1	1	3.56	0.36	9.86	<.0001
x_2	1	-1.64	0.15	-10.28	<.0001
x_3	1	0.33	0.17	1.88	0.0741

The value of R^2 of this model is 0.93. From Table 2, we see that F-statistic is highly significant, indicating that overall model fitting is good. R^2 is also very high. The fitted regression line is $\hat{y} = 8.19 + 3.56\,x_1 - 1.64\,x_2 + 0.33x_3$. The corresponding standard errors are given in the 4*th* column of Table 3. However, while testing the significance of the parameter estimates, we find that the intercept and the parameter for the variable x_3, *i.e.*, β_3 are not significant at 5% level of significance (probability values for these parameters are greater than 0.05). Since intercept term is not significant, one may interested to fit this model without intercept. The fitted model summary is given in Table 4 and Table 5.

Table 4: ANOVA with no intercept model

Source	Degrees of freedom	Sum of Square	Mean Square	F-value	Prob. > F
Model	3	76313	25438	7647.29	<.0001
Error	23	76.50	3.32		
Corrected Total	26	76389			

Table 5: Parameter Estimates with no intercept model

Variable	Degrees of freedom	Parameter Estimates	Standard Error	t-value	Prob. > |t|
x_1	1	4.02	0.07	54.36	<.0001
x_2	1	-1.53	0.13	-11.04	<.0001
x_3	1	0.51	0.11	4.36	0.0002

The R^2 value has been increased to 0.99. The dramatic change to notice here is that all parameter estimates are highly significant. The fitted model is $\hat{y} = 4.02\ x_1 - 1.53\,x_2 + 0.51x_3$. With this 'good' result one may be tempted to use a 'no intercept' model to report his/her findings. However, the situation from where the data is collected may demand a model with intercept. But intercept in that model is not significant. What to do? Actually our investigation starts from here. Many things essential for regression analysis have been ignored while fitting the model including statistical assumptions. One must check whether model assumptions required for analysis are satisfied or not before inferring from a data set. Once these assumptions are satisfied, we can go ahead for further analysis. If any one of these assumptions is violated, we have to some remedial measures to rectify problem. In the present case intercept term becomes non-significant may due to non fulfillment of some assumptions. In the next section we describe how to carry on a diagnostic check to see whether model assumptions are satisfied or not.

3 Diagnostics

As mentioned earlier when a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this section we discuss some simple methods for studying the appropriateness of a model.

Generally following departures may happen with a linear regression model.

(i) *The linearity of regression function.*

(ii) *The normal distribution of error terms.*

(iii) *The constancy of error variance.*

(iv) *The independency of error terms.*

(v) *Presence of one or a few outlier observations.*

(vi) *One or several important predictor variables have been omitted from the model.*

(vii) *Presence of multicollinearity.*

We now describe some tests to detect these departures.

3.1 Linearity of Regression Model

Whether a linear regression function is appropriate for the data being analyzed can be studied by plotting residuals against the predictor variables or equivalently against the fitted values. Figure 1 shows a prototype situation of the residual plot against *x* when a linear regression model is appropriate. In this plot the residuals fall within a horizontal band centred around 0, displaying no systematic tendencies to be positive and negative. Thus when the residuals scattered around zero, we say that linearity assumption is satisfied, otherwise not. Figure 1 presents a prototype plot of this situation.

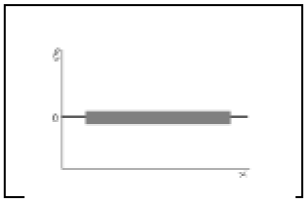


Figure 1: Prototype residual plot

From our data we plotted residuals against the fitted values. This plot is displayed in Figure 2.

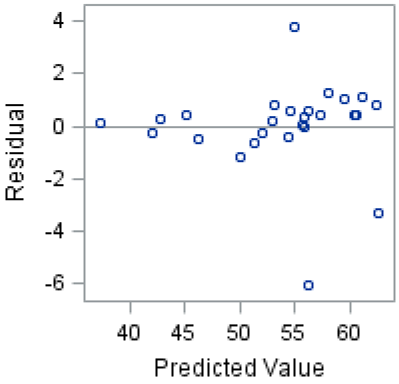


Figure 2: Residual plot of the data

From this plot it is evident that apart from a few points, most of the residuals are scattering around zero, indicating that the assumption of linearity of the regression function is satisfied.

3.2 Normality of Errors

Small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. There are many tests available for testing normality of errors. Here we discuss some of these tests in brief.

Comparison of frequencies: When the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 90% of the residuals fall between $\pm 1.645 \sqrt{MSE}$.

Correlation Test for Normality: A formal test for normality of the error terms can be conducted by calculating the coefficient of correlation between residuals r_i and their expected values under normality. A high value of the correlation coefficient is indicative of normality.

Kolmogorov-Smirnov test: The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given *n* ordered residuals $r_{(1)}, r_{(2)}, \ldots, r_{(n)}$, the ECDF is defined as

$$E_n = \frac{n(i)}{n},$$

where $n(i)$ is the number of points less than $r_{(i)}$ and the $r_{(i)}$ are ordered from smallest to largest value. This is a step function that increases by 1/*n* at the value of each ordered data point. Then the distance between the empirical distribution function and a normal cumulative distribution function is computed for each point. The K-S test is based on the maximum distance between these two distributions. An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test

The hypotheses tested under Kolmogorov-Smirnov test are

H_0: The data follow a specified distribution

H_1: The data do not follow the specified distribution

The Kolmogorov-Smirnov test statistic is then defined as

$$D = \max_{1 \le i \le n}\left(F(r_{(i)}) - \frac{i-1}{n},\ \frac{i}{n} - F(r_{(i)})\right) \tag{8}$$

where *F* is the theoretical cumulative distribution of the normal distribution. The hypothesis regarding the distributional form is rejected if the test statistic, *D*, is greater than the critical value obtained from a table which is available in the literature.

Anderson-Darling Test: The Anderson-Darling test is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov

(K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I, and logistic distributions.

The Anderson-Darling test is defined as:

H_0: The data follow a specified distribution.

H_1: The data do not follow the specified distribution

The Anderson-Darling test statistic is defined as $A^2 = -n - S$,

where, $S = \sum_{i=1}^{n} \frac{(2i-1)}{n} [\ln F(r_{(i)}) + \ln(1 - F(r_{(n+1-i)})]$ (9)

F is the cumulative distribution function of the specified distribution. Note that the $r_{(i)}$ are the *ordered* data (residuals in our case). The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas are available in literature for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value.

Some of the test-statistics available in the literature are applied to the present data. The results are given in Table 6. From the table, it is seen that all test-statistics values are highly significant. Thus null hypothesis is rejected and concluded that the error distribution is not normal. The theoretical normal curve is matched with the histogram obtained from the data. It is presented as Figure 3. Figure 3 also indicates that the distribution of error is not normal. All test like *t*- and *F*- are based on the assumption of the normal distribution of error. Since, the present data does not satisfy the assumption of normality, *t* and *F* values presented in Tables 2, 3, 4 and 5 are not reliable.

Table 6: Test For Normality

Name of Test	Teast-value	p-value
Shapiro-Wilk	0.761333	<0.0001
Kolmogorov-Smirnov	0.244031	<0.0100
Cramer-von Mises	0.43331	<0.0050
Anderson-Darling	2.357103	<0.0050

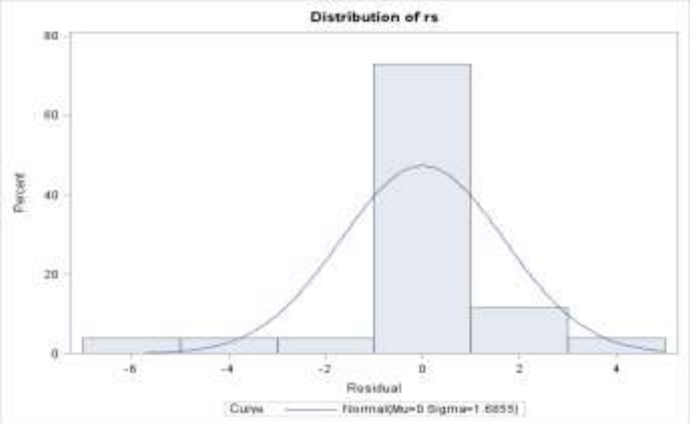


Figure 3: Normal curve along with the histogram of the data

3.3 Nonconstancy of Error Variance

Plots of residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. The prototype plot in Figure 1 exemplifies residual plots when error term variance is constant. However, there are many statistical tests available in the literature for testing constancy of error variance. We describe two such popular tests here.

Modified Levene Test: The test is based on the variability of the residuals. We divide all residuals in two groups. Let r_{i1} denotes the i^{th} residual for group 1 and r_{i2} denotes the i^{th} residual for group 2. Also we denote n_1 and n_2 to denote the sample sizes of the two groups, where: $n_1 + n_2 = n$. Further, we use $\tilde{r}_1$ and $\tilde{r}_2$ to denote the medians of the residuals in these groups. The modified Levene test uses the absolute deviations of the residuals around their median (in original Levene test mean was used in place of median), to be denoted by d_{i1} and d_{i2}:

$$\bar{d}_{i1} = |r_{i1} - \tilde{r}_1|, \quad \bar{d}_{i2} = |r_{i2} - \tilde{r}_2|.$$

The test used is t-test. The t-statistic is given as

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad , \tag{10}$$

where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2}, respectively, and the pooled variance s^2 is:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2} .$$

If the error terms have constant variance and n_1 and n_2 are not too small, t_L^* follows approximately the t distribution with $n-2$ degrees of freedom. Large absolute values of t_L^* indicate that the error terms do not have constant variance.

***White Test*:** The White test is a statistical test that establishes whether the residual variance of a variable in a regression model is constant, *i.e.*, homoscedastic. This test was proposed by Halbert White in 1980. This test is widely used test in practice. To test for constant variance of errors one has to carry out an auxiliary regression analysis. This regression is carried out by taking the squared residuals from the original regression model as dependent variable and a set of regressor variables, which contains the original regressors, the cross-products of the regressors and the squared regressors. One then compute R^2 from this fitting. The test statistic is the product of the R^2 value and sample size

$$LM = n.R^2$$

(12)

This follows a chi-square distribution, with degrees of freedom equal to the number of estimated parameters (in the auxiliary regression) minus one.

For the present data, White test and an improvement over this test Breusch-Pagan test were carried out. Results are presented in Table 7. From this Table, it is evident that the null hypothesis of equality of variance is accepted. Thus the present data is homoscedastic.

Table 7: Test for homoscedasticity

Test	Statistic	Degrees of freedom	Pr > ChiSq
White's Test	26.04	25	0.4056
Breusch-Pagan	3.90	9	0.9181

3.4 Independence of Error Terms

A run test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specially designed for lack of randomness in least squares residuals, is the Durbin-Watson test.

Durbin-Watson test: The Durbin-Watson test assumes the first order autoregressive error models. The test consists of determining whether or not the autocorrelation coefficient (ρ, say) is zero. The hypotheses for this test are:

$H_0 : \rho = 0$

$H_0 : \rho > 0$

The Durbin-Watson test statistic D is obtained by calculating the ordinary residuals r_t, and then calculating the statistic:

$$D = \frac{\sum_{t=2}^{n} (r_t - r_{t-1})^2}{n \sum_{t=1} r_t^2} \qquad (13)$$

Exact critical values are difficult to obtain, but Durbin-Watson have obtained lower and upper bound d_L and d_U such that a value of *D* outside these bounds leads to a definite decision. The decision rule for testing between the alternatives is:

if $D > d_U$, conclude H_0

if $D < d_L$, conclude H_1

if $d_L \le D \le d_U$, test is inconclusive.

Small value of *D* lead to the conclusion that ρ >0.

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographical areas, it is good idea to prepare a sequence plot of the residuals. The purpose of plotting the residuals against time or some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence.

3.5 Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The purpose of this additional analysis is to determine whether there are any key variables that could provide important additional descriptive and predictive power to the model. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable. If a particular predictor variable vary systematically with the residual, we say that this variable is important and refit the model using this variable also as a predictor variable.

3.6 Tests for Outliers

As a rough definition, we refer to outliers as those observations for which the inputs are reasonable but the response is abnormally large or small as compared to other cases with similar inputs. Extreme cases are those for which the input value is, in some cases, far from rest of the data. Such cases are also referred to as high leverage points. In either case, these observations may greatly influence the least squares estimates. Detecting these observations, therefore, is very important. Here we discuss some important test-statistics for detecting outliers in linear regression model.

Elements of Hat Matrix (h_{ii}) **:** The Hat matrix is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, where **X** is obtained using all explanatory variables. The larger values reflect data points are outliers.

WSSD$_i$: *WSSD$_i$* is an important statistic to locate points that are remote in *x*-space. *WSSD$_i$* measures the weighted sum of squared distance of the i^{th} point from the center of the data. Generally if the *WSSD$_i$* values progress smoothly from small to large, there are probably no extremely remote points. However, if there is a sudden jump in the magnitude of *WSSD$_i$*, this often indicates that one or more extreme points are present.

Cook's D_i: *Cook's D_i* is designed to measure the shift in $\hat{y}$ when i^{th} obsevation is not used in the estimation of parameters. D_i follows approximately $F_{(p,n-p-1)}$(1-α). Lower 10% point of this distribution is taken as a reasonable cut off (more conservative users suggest the 50% point). The cut off for D_i can be taken as $\frac{4}{n}$, where n is the total number of observations.

DFFITS_i : *DFFIT* is used to measure difference in i^{th} component of $\left(\hat{y}-\hat{y}_{(i)}\right)$, where is obtained after deleting the i^{th} data point. It is suggested that $DFFITS_i \geq 2\left(\frac{p+1}{n}\right)^{1/2}$ may be used to flag off influential observations.

DFBETAS$_{j(i)}$: Cook's D_i reveals the impact of i^{th} observation on the entire vector of the estimated regression coefficients. The influential observations for individual regression coefficient are identified by $DFBETAS_{j(i)}, j = 1,2,...,p$, where each $DFBETAS_{j(i)}$ is the standardized change in β_j when the i^{th} observation is deleted.

COVRATIO$_i$:The impact of the i^{th} observation on variance-covariance matrix of the estimated regression coefficients is measured by the ratio of the determinants of the two variance-covariance matrices, one is obtained with full data and the other is obtained after deleting the i^{th} data point. Thus, COVRATIO reflects the impact of the i^{th} observation on the precision of the estimates of the regression coefficients. Values near 1 indicate that the i^{th} observation has little effect on the precision of the estimates. A value of COVRATIO greater than 1 indicates that the deletion of the i^{th} observation decreases the precision of the estimates; a ratio less than 1 indicates that the deletion of the observation increases the precision of the estimates. Influential points are indicated by $\left|COVRATIO_i - 1\right| > \frac{3(p+1)}{n}$.

FVARATIO$_i$: The statistic detects change in variance of $\hat{y}_i$ when an observation is deleted. A value near 1 indicates that the i^{th} observation has negligible effect on variance of y_i. A value greater than 1 indicates that deletion of the *i^th^* observation decreases the precision of the estimates, a value less than one increases the precision of the estimates.

Table 8: Indicators of Influential Observations

Case	$\boldsymbol{h_{ii}}$	$\boldsymbol{D_i}$	$\boldsymbol{WSSD_i}$	***Cov Ratio***	***Dffits***	***DFBETAS***			
						Intercept	β_0	β_1	β_2
1	0.215	0.005	39	1.512	0.148	0.056	-0.053	-0.006	0.006
2	0.093	0.013	12	1.203	0.232	0.062	-0.042	-0.042	-0.050
3	0.048	0.001	1	1.254	0.047	-0.005	0.010	-0.008	-0.007
4	0.042	0.000	1	1.257	0.005	0.000	0.000	-0.001	0.000
5	0.053	0.000	3	1.267	-0.033	-0.001	-0.001	-0.006	0.006
6	0.155	0.017	20	1.331	0.258	-0.095	0.132	-0.042	-0.050
7	0.081	0.001	7	1.299	0.068	-0.005	0.015	-0.036	-0.005
8	0.301	0.001	41	1.721	0.057	0.027	-0.034	0.026	-0.006
9	0.155	0.003	18	1.408	0.109	-0.030	0.048	-0.035	-0.031
10	0.147	0.005	23	1.380	0.144	0.058	-0.058	-0.041	0.016
11	**0.173**	**0.214**	14	**0.639**	**-1.004**	**-0.154**	-0.045	**0.776**	**0.525**
12	0.053	0.001	3	1.260	-0.054	-0.017	0.014	0.014	0.000
13	0.163	0.001	24	1.435	0.051	0.017	-0.19	-0.004	0.013
14	0.175	0.001	23	1.452	-0.074	-0.026	0.031	-0.35	0.015
15	0.122	0.007	15	1.315	0.175	-0.008	0.033	-0.105	0.002
16	0.177	0.005	26	1.441	-0.134	-0.014	0.014	-0.044	0.047
17	**0.041**	0.048	**0**	**0.496**	**0.482**	0.061	**-0.17**	**-0.107**	-0.046
18	0.114	**0.412**	**8**	**0.410**	**-1.945**	**0.362**	**-0.308**	**-0.220**	**-1.177**
19	0.160	0.025	24	1.301	-0.341	0.031	-0.045	-0.080	0.094
20	0.114	0.014	11	1.252	0.236	-0.055	0.097	-0.105	-0.051
21	0.119	0.003	12	1.350	0.095	0.054	-0.061	0.024	-0.018
22	0.055	0.003	3	1.228	0.108	0.052	-0.048	-0.028	-0.020
23	0.059	0.000	3	1.279	-0.008	0.001	-0.002	0.001	0.002
24	**0.927**	**4.409**	**19**	**12.715**	**4.230**	**-3.642**	**3.276**	**3.180**	**3.934**
25	0.159	0.001	19	1.426	0.069	0.031	-0.039	0.029	-0.003
26	0.101	0.004	11	1.309	-0.117	0.000	-0.007	-0.016	0.043

For the present data, some of these statistics are used to detect outliers, if any. The results are presented in Table 8. From this Table, we see that observation Numbers 11, 17, 18 and 24 stand out with respect to some of the statistics. These observations are tested with cut-off values and found to be outliers.

3.7 Tests for Multicollinearity

The use and interpretation of a multiple regression model depends implicitly on the assumption that the explanatory variables are not strongly interrelated. In most regression applications the explanatory variables are not orthogonal. In some situations the explanatory variables are so strongly interrelated that the regression results are ambiguous. Typically, it is impossible to estimate the unique effects of individual variables in the regression equation. The estimated values of the coefficients are very sensitive to slight changes in the data and to the addition or deletion of variables in the equation. The regression coefficients have large sampling errors which affect both inference and forecasting. The condition of severe non-orthogonality is also referred to as the problem of *multicollinearity*. Multicollinearity also tends to produce least squares estimates β_j that are too large in absolute value.

Detection of Multicollinearity: Let $R = (r_{ij})$ and $R^{-1} = (r^{ij})$ denote simple correlation matrix and its inverse. Let $\lambda_i, i = 1,2,...,p$ $(\lambda_p \le \lambda_{p-1} \le ....\lambda_1)$ denote the eigen values of *R*. The following are common indicators of relationships among independent variables.

(i) Simple pair-wise correlations $|r_{ij}| = 1$

(ii) The squared multiple correlation coefficients $R_i^2 = 1 - \frac{1}{r^{ii}} > 0.9$, where R_i^2 denote the squared multiple correlation coefficients for the regression of x_I on the remaining *x* variables.

(iii) The variance inflation factors, $VIF_i = r^{ii} > 10$ and

(iv) eigen values, $\lambda_i = 0$.

The first of these indicators, the simple correlation coefficients between pairs of independent variables r_{ij}, may detect a simple relationship between x_i and x_j. Thus $|r_{ij}| = 1$ implies that the *i*th and *j*th variables are nearly proportional.

The second set of indicators, R_i^2, the squared multiple correlation coefficient for the regression of x_i on the remaining *x* variables indicates the degree to which x_i is explained by a linear combination of all of the other input variables.

The third set of indicators, the diagonal elements of the inverse matrix, which have been labeled as the Variance Inflation Factors, VIF_i. The term arises by noting that with standardized data (mean zero and unit sum of squares), the variance of the least squares estimate of the ith coefficient is proportional to r^{ii}, $VIF_i > 10$ is probably based on the simple relation between R_i and VIF_i. That is $VIF_i > 10$ corresponds to $R_i^2 > 0.9$.

Sometimes condition numbers are used. It equals the square root of the largest eigenvalue (λ_1) divided by the smallest eigenvalue (λ_p), *i.e.*,

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}\,.$$

When there is no collinearity at all, the eigenvalues and condition number will all equal one. As collinearity increases, eigenvalues will be both greater and smaller than 1 (eigenvalues close to zero indicate a multicollinearity problem), and the condition number will increase. An informal rule of thumb is that if the condition number is 15, multicollinearity is a concern; if it is greater than 30 multicollinearity is a very serious concern.

Table 9: Collinearity Diagnostics

Variable	***VIF***	***Eigenvalue***	***Condition Index***
Intercept	0	3.38175	1.00000
x1	1.76883	0.53975	2.50308
x2	1.86011	0.07670	6.64020
x3	2.82000	0.00180	43.31982

For the present data these are worked out and presented in Table 9. On the basis of these values we conclude that variable x_3 is creating the problem of multicollinearity.

Thus we, see that in the present data, there are four outliers and one variable which causes multicollinearity. As a remedial measures we delete these four observations and dropped variable x_3. The results are presented in Table 10 and Table 11. The dramatic change now can be noticed. All parameter estimates are now highly significant. *F*-statistic is highly significant as well as the model fit has a high R^2 (0.99) value.

Table 10: ANOVA with cleaned data

Source	**Degrees of freedom**	**Sum of Square**	**Mean Square**	**F-value**	**Prob. > F**
Model	2	1048.11	524.05	2221.21	<.0001
Error	19	4.48	0.23		
Corrected Total	21	1052.60			

Table 11: Parameter Estimates of model with cleaned data

Variable	Degrees of freedom	Parameter Estimates	Standard Error	t-value	Prob. > |t|
Intercept	1	19.45	1.16	16.74	<0.0001
x_1	1	3.03	0.07	38.63	<0.0001
x_2	1	-1.99	0.03	-59.09	<0.0001

However, our data was non-normal. We have not taken any remedial measure for it. But after removal of outliers and rectifying the problem of multicollinearity, our data may become normal. We, therefore, again tested for normality on the rectified data. The results are presented in Table 13. It is now seen that the errors of the reduced data is normal. Figure 4 which shows the histogram with normal curve also confirms this fact.

Table 13: Test for normality of cleaned data

Name of Test	Teast-value	p-value
Shapiro-Wilk	0.980664	0.9261
Kolmogorov-Smirnov	0.136376	>0.1500
Cramer-von Mises	0.03889	>0.2500
Anderson-Darling	0.223155	>0.2500

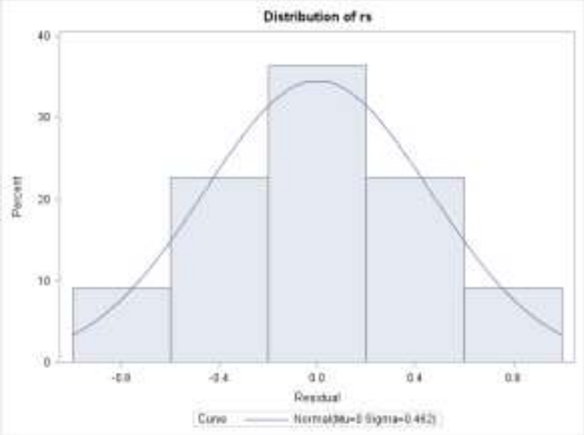


Figure 4: Normal curve and Histogram of cleaned data

The remedial measures taken above are not the only measures available. There are other measures which can be applied when any of the above assumptions is violated. We now discuss these remedial measures in brief in the next Section.

4. Remedial Measures

If the regression model is not appropriate for a data set, there are two basic choices:

(i) *Abandon regression model and develop and use a more appropriate model.*

(ii) *Employ some transformation on the data so that regression model is appropriate for the transformed data.*

Each approach has advantages and disadvantages. The first approach may entail a more complex model that could yield better insights, but may also lead to more complex procedure for estimating the parameters. Successful use of transformations, on the other hand, leads to relatively simple methods of estimation and may involve fewer parameters than a complex model, an advantage when the sample size is small. Yet transformation may obscure the fundamental interconnections between the variables.

Nonlinearity of Regression Function

When the regression function is not linear, a direct approach is to modify regression model. For example, we can modify the simple regression model (1) by altering the nature of the regression function. For instance, a quadratic regression function might be used.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

or an exponential regression function:

$$y_i = \gamma_0 \gamma_1^{x_i} + \varepsilon_i \,.$$

When the nature of the regression function is not known, exploratory analysis that does not require specifying a particular type of function is often useful.

Nonconstancy of Error Variance

When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the method to allow for this and use the method of weighted least squares to obtain the estimates of the parameters.

Transformation is another way in stabilizing the variance. We first consider transformation for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformation on *x* should be attempted. The reason why transformation on *y* may not be desirable here is that a transformation on *y*, such as $y' = \sqrt{y}$, may materially change the shape of the distribution and may lead to substantially differing error term variance.

Following transformations are generally applied for stabilizing variance.
(1) when the error variance is rapidly increasing $y' = \log_{10} y$ or $y' = \sqrt{y}$
(2) when the error variance is slowly increasing, $y' = y^2$ or $y' = Exp(y)$
(3) when the error variance is decreasing, $y' = 1/y$ or $y' = Exp(-y)$.

Box - Cox Transformations: It is difficult to determine, which transformation of *y* is most appropriate for correcting skewness of the distributions of error terms, unequal error variance, and nonlinearity of the regression function. The Box-Cox transformation automatically identifies a transformation from the family of power transformations on *y*. The family of power transformations is of the form: $y' = y^{\lambda}$, where is a parameter to be determined from the data. Using standard computer programme it can be determined easily.

Nonindependence of Error Terms

When the error terms are correlated, a direct approach is to work with a model that calls for error terms. A simple remedial transformation that is often helpful is to work with first differences.

Nonnormality of Error terms

Lack of normality and non-constant error variance frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore, desirable that the transformation for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present.

Omission of Important Variables

When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model.

Outlying Observations

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents in error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstances. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. Robust Regression falls under such methods.

Multicollinearity

(i) **Collection of additional data:** Collecting additional data has been suggested as one of the methods of combating multicollinearity. The additional data should be collected in a manner designed to break up the multicollinearity in the existing data.

(ii) **Model respecification:** Multicollinearity is often caused by the choice of model, such as when two highly correlated regressors are used in the regression equation. In these situations some respecification of the regression equation may lessen the impact of multicollinearity. One approach to respecification is to redefine the regressors. For example, if x_1, x_2 and x_3 are nearly linearly dependent it may be possible to find some function such as $x = (x_1+x_2)/x_3$ or $x = x_1x_2x_3$ that preserves the information content in the original regressors but reduces the multicollinearity.

(iii) **Ridge Regression:** When method of least squares is used, parameter estimates are unbiased. A number of procedures have been developed for obtaining biased estimators of regression coefficients to tackle the problem of multicollinearity. One of these procedures is

ridge regression. The ridge estimators are found by solving a slightly modified version of the normal equations. Each of the diagonal elements of $\mathbf{X'X}$ matrix are added a small quantity.

Note: *This note is prepared on the basis of materials taken from the references cited below and statistical analysis of the example is done through SAS software available at Indian Agricultural Statistics Research Institute, New Delhi***.**

<u>Some Selected References</u>

Belsley, D. A., Kuh, E. and Welsch, R. E. (2004). *Regression Diagnostics – Identifying Influential Data and Sources of Collinearity*, New York: Wiley.

Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*, New York: Wiley.

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*, New York: John Wiley & Sons.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, New York: Wiley Eastern Ltd.

Kleinbaum, D. G. and Kupper, L. L. (1978). *Applied Regression Analysis and Other Multivariate Methods*, Massachusetts: Duxbury Press

Montgomery, D. C., Peck, E. and Vining, G. (2003). *Introduction to Linear Regression Analysis*, 3rd Edition, New York: John Wiley and Sons.

Modeling the Growth of *Lactococcus lactis*.

Sunita Singh
ICAR-IARI, New Delhi
Sunita.Singh@icar.gov.in

Microbial growth kinetics can be determined by different models. A mathematical model is an expression of a defined equation, or a set of equations, that attempts to explain instances of reality in a simplified manner, utilizing only a system's most pertinent properties [Pérez-Rodríguez and Valero, 2013].

Models can have three levels: 1) Primary level models that describe changes in microbial numbers with time, 2) Secondary level models that show how the parameters of the primary model vary with environmental conditions, and 3) Tertiary level that combine the first two types of models with user-friendly application software or expert systems to calculate microbial behavior under the specified conditions (Whiting, 1995). Primary models include time-to-growth, Gompertz function, exponential growth rate, and inactivation/survival models. Commonly used secondary models are response surface equations and the square root and Arrhenius relationships. Such models are used to describe the behavior of microorganisms under different physical or chemical conditions such as temperature, pH, and water activity. These models allow the prediction of microbial growth, safety or shelf life of products, the detection of critical parts of the production and distribution process, and the optimization of production and distribution chains (Zwiettering, *et al*., 1990). Thus to build these models, firstly growth has to be measured using models. These models include: (1) Monod, (2) Gompertz, (3) Contois, (4) Baranyi-Roberts, (5) logistic, 6) Richards, 7) Schnute, and 8) Stannard equations/functions [Monod, 1949; Zwiettering, *et al*., 1990; Contois, 1959; Grijspeerdt and Vanrolleghem, 1999). The usefulness of any empirical equation of such a model is enhanced, if the constants easily yield information, on direct biological interest (Richards, 1959).

Such microbial models can be used to compare and describe a bacterial growth curve (Zwiettering, et al., 1990; Esser et al., 2015). On the other hand microbial death has also been modeled in predictive microbiology applications for different applications (challenge test, evaluation of microbiological shelf life, prediction of the microbiological hazards connected with foods, etc (Bevilacqua *et al*., 2015). Having said this, where predictive modeling has applications focused on mathematical models for microbial inactivation, the primary models used in those studies are still necessary for information on growth, for predictive applications, where inhibition and death kinetics is the major concern. How well they fit to predict experimental growth data, is always the target.

For increasing overall predictive accuracy, purpose of modeling growth (batch culturing, continuous or fed-batch culture) or predictive modeling as in microbial inactivation, must be clearly defined, to help improve models.

To determine growth characteristics of microbes in absence of software or other appropriate methods, the use of subjective graphical interpretations of linearized logarithmic data are also most commonly used. In comparison to graphical determinations, modified Maclaurin series have also been evaluated for microbial growth kinetics (Talkington, et al., 2013) in bacteria.

In this lecture we are going to focus on primary microbial growth modeling. This can be applied to obtain information on the specific growth rate to exploit the harvest of useful microbial metabolite(s) nisin (an antimicrobial compound) excreted by *Lactococcus lactis*, a lactic acid bacteria (LAB). In describing the study (Singh *et al*., 2015), a batch fermenter with a closed habitat was used that showeda typical growth of bacteria with four growth stages namely, a “lag phase,” “exponential growth phase,” a “stationary phase,” and a“mortality phase” (McKellar and Lu, 2004). The specific growth rates of bacterial populations are a function of their population density (Bail, 1929). On the other hand bacterial concentrations have also been used to interpret growth in various studies.

Initially an aim is set to evaluate differences (if any) in the 3 different sigmoidal growth functions used to model growth of *L. lactis*. As the bacteria grows exponentially, the sigmoidal functions (Gompertz , logistic) can be compared to Richards function that has a fourth parameter known to describe shape of growth curve.

Using the SPSS software to fit data, the RSS (residual sum of squares) values are calculated for the three functions. The three-parameter functions can be statistically compared to the four-parameter function by the F test, to discriminate among the three-parameter functions (Logistic/Gompertz) from four-parameter function (Richards). The fitted data are analyzed and used in calculating f_d values that are tested against F (Table values). The residual sum of squares RSS_1 (of fourth parameter Richards function), can be used as a measuring error (Zwiettering *et al*., 1990) to check the acceptability of either of the three-parameter sigmoidal functions, in order to model growth of *L. lactis*.

$$\text{Logistic} \quad y = \frac{a}{[1+\exp(b-ct)]}$$

$$\text{Gompertz} \quad y = a * \exp[-\exp(b-ct)]$$

$$\text{Richards} \quad y = a\{1+b*\exp[c*(d-t)]\}^{-1/b}$$

An algorithm (Marquardt), is used to fit growth data to these three nonlinear equations/functions [Logistic; Gompertz & Richards] as it removes the divergence of successive iterates while fitting data points, by the nearest neighborhood method, assuming local linearity at each iteration (as in Taylor series) [Marquardt, 1963]. This algorithm, closed in on the converged values rapidly after the vicinity of the converged values have been reached.

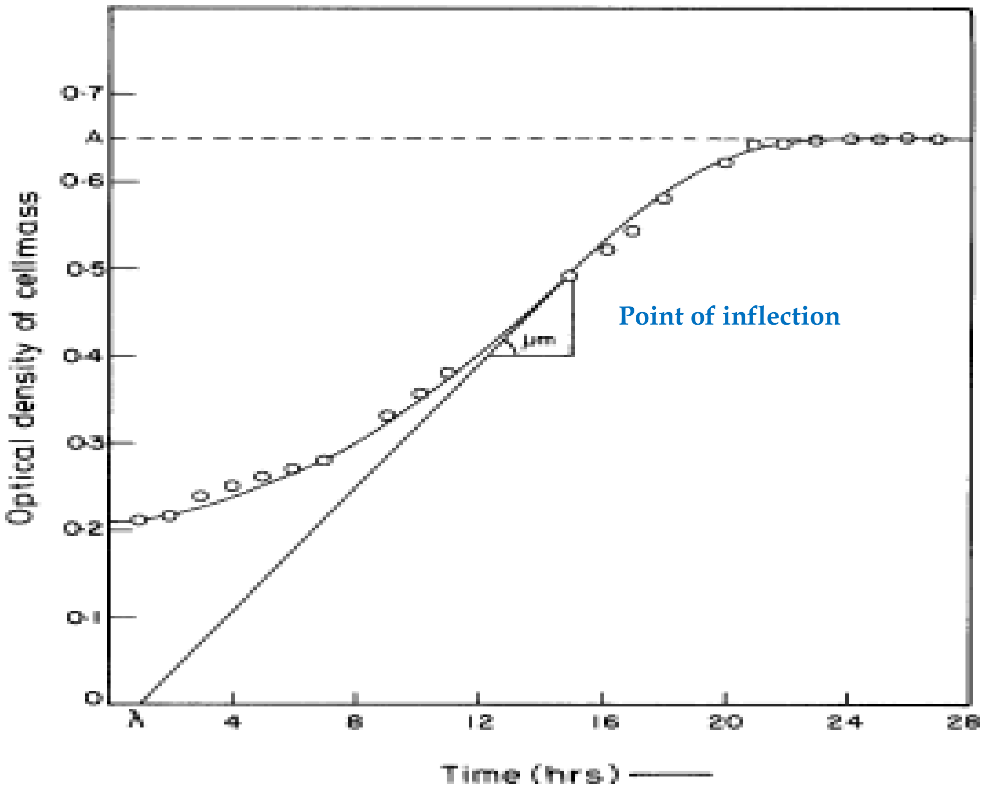
The steps towards curve fitting:

- Nonlinear regression using a Marquardt algorithm wherein the iteration method was by least square estimation.
- Software SPSS 17.1 version was used to then fit data of the curve for a best fit curve.

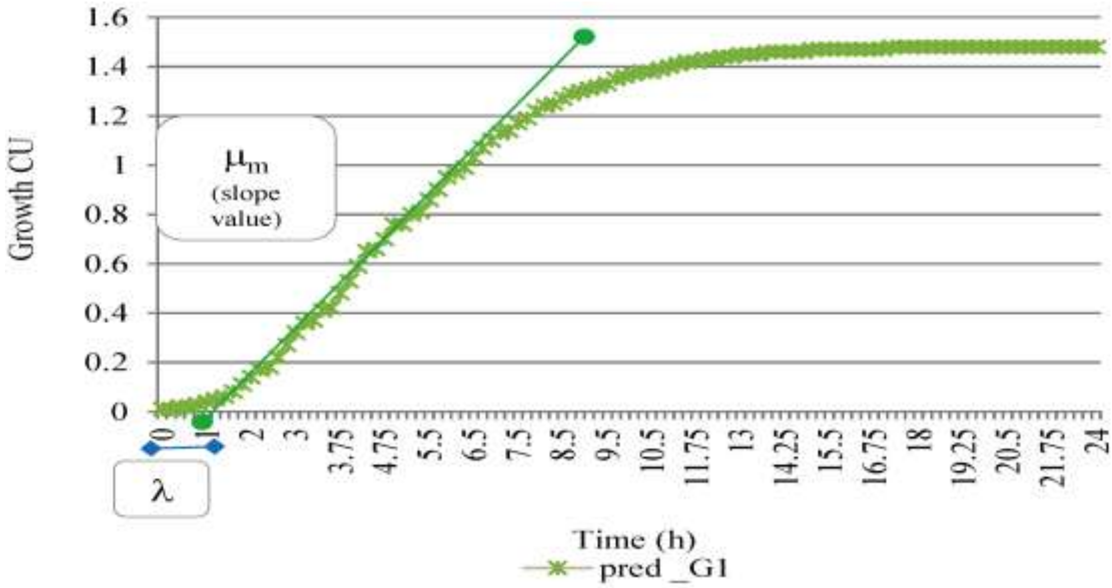
The LAB, *Streptococcus lactis* NCIM 2114 produced nisin as a primary metabolite. Being a primary metabolite, the rate of nisin formed depends on its growth rate.

Thus aiming to simplify the kinetics and describe the entire set of data with a growth model to estimate A, μ_m and λ from the model, the function that best described growth curve (here Gompertz), was used to obtain mathematical parameters (a, b, c) rather than parameters with a biological meaning (A, μ_{max}, and λ). Differentiating this function, to obtain simple equations in terms of the mathematical parameters (Zwiettering, et al., 1990) helped to obtain simple equations to calculate specific growth rate of the bacteria (Singh et al., 2015).

The maximum specific growth rate (μ_{max}) during exponential growth of bacteria also graphically corresponds to the slope of the log-linear part of the exponential growth curve and remains constant in that phase (Fig 1). The expression for the maximum specific growth rate was derived by calculated by the first derivative of function as the tangent on the inflection point (Zwiettering, *et al.*, 1990). The second derivative of Gompertz function equals zero at the μ_{max} and rate of change was maximum. In calculating microbial kinetics, the value of the inflection point is determined to pinpoint the shift from exponential increase to exponential decrease (Talkington *et al.*, 2013) of growth. The inflection point is thus the maximum value of the signal prior to the beginning of the exponential decrease.



a



b

Fig 1. Microbial growth. [b)Source: Singh *et al*., 2015].

The lag time λ, needed for the population to adjust to its new environment, is marked at the X-axis intercept of the tangent on the inflection point (Fig 1). This point determines the point at which a tangent, drawn over the exponential phase, intercepts the horizontal axis to determine the lag time of the growth phase. The asymptote [$A = \ln(N_\infty/N_0)$] is the maximal value of growth reached (Zwiettering, *et al*., 1990).

The mathematical parameter values were then obtained by using the equations :

$$c = \frac{\mu_{\max} e}{a}$$

$$b = \frac{\mu_{\max} e}{a}\lambda + 1$$

The study of the growth curve (characteristically a sigmoid shape growth), from lag to exponential phase can benefit to estimate lag period and the specific growth rate with which the microorganism can be handled or manipulated under a known set of environmental conditions.

Bibliography:

Bail, 0. (1929). Ergebnisse experimenteller Populationsforschung. *Z. Immun-Forsch*. 60: 1.

Bevilacqua Antonio, Speranza Barbara, Sinigaglia Milena and Corbo Maria Rosaria (2015). A Focus on the Death Kinetics in Predictive Microbiology: Benefits and

Limits of the Most Important Models and Some Tools Dealing With their Application in Foods. *Foods*. 4: 565-580.

Contois DE. (1959). Kinetics of Bacterial Growth: Relationship between Population Density and Specific Growth Rate of Continuous Cultures. *J. Gen. Microbiol*. 21(1) :40-50.

Esser Daniel S., Leveau Johan HJ. and Meyer Katrin M. (2015). Modeling Microbial Growth and Dynamics. *Appl Microbiol Biotechnol*. DOI 10.1007/s00253-015-6877-6.

Grijspeerdt K and Vanrolleghem P. (1999). Estimating the Parameters of the Baranyi Model for Bacterial Growth. *Food Microbiol*. 16(6): 593-605.

Marquardt DW. (1963). An algorithm for Least-Squares Estimation of Nonlinear Parameters. *J. Soc. Ind. Appl. Math*. 11:431-441.

McKellar R. and Lu X. (Eds.) (2004). Modeling Microbial Responses on Foods. CRC Press, Boca Raton, FL.

Monod J. (1949). The Growth of Bacterial Cultures. *Ann. Rev. Microbiol*. 3: 371-394.

Peleg Micha and Corradini Maria G. (2011). Microbial Growth Curves: What the Models Tell us and What they Cannot. *Critical Rev. Food Sci. Nutr*. 51(10): 917-945.

Pérez-Rodríguez, Valero FA. (2013). Predictive Microbiology in Foods. In “*Predictive Microbiology in Foods*” Springer: New York, NY, USA. pp. 1–10.

Sunita Singh, Kamalesh N. Singh, Siva Mandjiny and Leonard Holmes (2015). Modeling the Growth of *Lactococcus lactis* NCIM 2114 Under Differently Aerated and Agitated Conditions in Broth Medium. *Fermentation*. *1*(1), 86-97.

Talkington Anne M., Inman III Floyd L. and Holmes Leonard D. 2013. A Novel Method for Determining Microbial Kinetics. *J. Life Sci.* 7(8): 787-790.

Whiting RC (1995). Microbial Modeling in Foods. Crit Rev Food Sci Nutr. 35(6): 464-94.

Zwietering MH, Jongenburger I, Rombouts FM, Riet KV. (1990). Modeling of the Bacterial Growth Curve. *Appl. Env. Microbiol*. 56(6): 1875-1881.

Hybrid Time Series Models

Wasi Alam and Santosha Rathod
ICAR-IASRI, New Delhi

In a developing country like India, food security means making available minimum quantity of food grains to the entire population. Despite the fact that India has made a satisfactory achievement in food grains production, its population growth has nullified the benefits of production. The FAO forecasts that global food production will need to increase by over 40% by 2030 and 70% by 2050 (FAO, 2009). Among food grains, rice is the most important crop of the developing world and the staple food for more than 60% of the Indian population. In India, the annual compounded growth rate of rice production has declined from 3.55 per cent during 1981-90 to 1.74 per cent during 1991-2000. Projection of rice demand/supply by 2030 mentioned in vision 2030 of Central Rice Research Institute (Adhya *et al*., 2011) has been computed on the basis of fixed historical growth rate rather than time series approach. Forecasting the future demand/supply of crop production to meet the need of corresponding future growing population is a major concern for policy planners. In order to get more reliable future crop production forecast, we need more precise time series forecast. Traditionally, classical ARIMA model (Box *et al.*, 2007; Cogger, 1988; Clements, 2003) has been widely used for short term time series forecasting. In ARIMA approach, the future value of a variable is assumed to be a linear function of several past observations and random errors. Classical ARIMA models are typically well-suited for short-term forecasts, but not for long term forecasts due to the convergence of the autoregressive part of the model to the mean of the time series. Moreover, this approach does not explain the nonlinear component of residuals obtained through ARIMA model. Here, we can improve the performance of ARIMA through the approach of Zhang (2003) in first instance and the improved forecast values can be used for long term forecast through the proposed technique. Rice is a rainfed crop and due to climate change rice yield may be tremendously influenced by weather variables. Hence, it is quite interesting to assess the impact of time series weather variables on yield and consequently rice production too. The ARIMAX model is a generalization of ARIMA model which is capable of incorporating an external input variable. Hyndman (2010) preferred to call ARIMAX as regression with ARIMA errors.

Univariate linear time series model

The univariate linear time series model (ARIMA)was proposed by Box-Jenkins (1970).In ARIMA, the future value of a variable is assumed to be a linear function of several past observations and random errors. Because of its relative simplicity in understanding and implementation, it has been the main research focuses and applied tools during the past few decades. Most time series can be described by Autoregressive Moving Average (ARMA) model. The stationary series Y_t is said to be ARMA(p, q) if

$$Y_t = \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where ε_t is white noise and there is no common factor between autoregressive polynomial, $(1-\phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)$ and moving average polynomial $(1+\theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)$, where L is a lag operator. Also, these polynomials can be represented by $\phi(L)$ and $\theta(L)$, respectively. If the series is not stationary then differencing is required to make the series stationary and the autoregressive integrated moving average (ARIMA) model is implemented and the series is called as ARIMA(p, d, q) if

$$\phi(L)(1-L)^d Y_t = \theta(L)\varepsilon_t$$

where d is the dth difference operator.

The stationarity of the series is important otherwise non-stationary series can strongly influence its behaviour and properties - e.g. persistence of shocks will be infinite for non-stationary series. If the variables in the regression model are not stationary, then it can be proved that the standard assumptions for asymptotic analysis will not be valid. In other words, the usual “t-ratios” will not follow a t-distribution, so we cannot validly undertake hypothesis tests about the regression parameters.

The Box–Jenkins methodology includes three iterative steps of model identification, parameter estimation and diagnostic checking. At identification stage, based on autocorrelation patterns (ACF or PACF) we identify one or several potential models for the given time series. Data transformation is often needed to make the time series stationary. Stationarity is a necessary condition in building an ARIMA model that is useful for forecasting. A stationary time series has the property that its statistical characteristics such as the mean and the autocorrelation structure are constant over time. When the observed time series presents trend and heteroscedasticity, differencing and power transformation are often applied to the data to remove the trend and stabilize the variance before an ARIMA model can be fitted. Once a tentative model is specified, model parameters are estimated such that an overall measure of errors is minimized via nonlinear optimization procedure.At the diagnostic checking stage, white noise test for the residuals of the tentatively identified candidate model is tested through many diagnostic statistics and plots of the residuals. If residuals are not white noise, we again select candidate model and repeat the same unless we get valid model.

ARIMAX Model

The ARMAX model is a generalization of ARMA model which is capable of incorporating an external input variable. ARIMA model is extended into ARIMA model with exogenous variable X, called ARIMAX (p, d, q). Let the time series be denoted by $\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_\mathrm{n}$ and we assume the series to be stationary, hence, we only consider ARMA model. First, we define an ARMA(p, q) model with no covariates:

$$y_t = \phi_1 y_{t\text{-}1} + \phi_2 y_{t\text{-}2} + ... + \phi_p y_{t\text{-}p} \text{ - } \theta_1 \varepsilon_{t\text{-}1} \text{ - } \theta_2 \varepsilon_{t\text{-}2} \text{ - } ... \text{ - } \theta_q \varepsilon_{t\text{-}q} + \varepsilon_t$$

Where ε_t is a white noise process (i.e. identically independently distributed with mean=0).

An ARMAX model simply adds in the covariate on the right hand side:

$$y_t = \beta x_t + \phi_1 y_{t\text{-}1} + \phi_2 y_{t\text{-}2} + ... + \phi_p y_{t\text{-}p} \text{ - } \theta_1 \varepsilon_{t\text{-}1} \text{ - } \theta_2 \varepsilon_{t\text{-}2} \text{ - } ... \text{ - } \theta_q \varepsilon_{t\text{-}q} + \varepsilon_t$$

Where x_t is a covariate at time t and β is its coefficient.

If we write the model using backshift operators, the ARMAX model is given by

$$\phi(B) y_t = \beta x_t + \theta(B) \varepsilon_t$$

or $y_t = \dfrac{\beta}{\phi(B)} x_t + \dfrac{\theta(B)}{\phi(B)} \varepsilon_t$,

where $\phi(B) = 1 - \phi_1 B - \ ... \ - \phi_p B^p$ and where $\theta(B) = 1 - \theta_1 B - \ ... \ - \theta_q B^q$

We note that autoregressive coefficients get mixed up with both the covariates and the error term. Hyndman (2010) preferred it to call regression with ARMA errors. Regression models with ARMA errors is defined as

$$y_t = \beta x_t + \eta_t$$
$$\eta_t = \phi_1 y_{t\text{-}1} + \phi_2 y_{t\text{-}2} + \ldots + \phi_p y_{t\text{-}p} - \theta_1 \varepsilon_{t\text{-}1} - \theta_2 \varepsilon_{t\text{-}2} - \ldots - \theta_q \varepsilon_{t\text{-}q} + \varepsilon_t$$

In this case, the regression coefficient has its usual interpretation. Using backshift operators, this model can be written as

$$y_t = \beta x_t + \frac{\theta(B)}{\phi(B)} \varepsilon_t$$

when the data is non-stationary, for the ARIMA errors, we simply replace $\phi(B)$ with $\nabla^d \phi(B)$ where $\nabla = 1 - B$ denotes the differencing operator. Differencing of y_t and x_t is required before fitting the model with ARMA errors. Hence, differencing all variables is necessary because estimation of a model with non-stationary errors is not consistent and can lead to spurious regression. The first step in building an ARMAX model consists of identifying a suitable ARMA model for the endogenous variable. The ARMAX model concept requires to test for stationarity of exogenous variable before modeling. Nonlinear least square estimation procedure is employed to estimate the parameters of ARMAX model. In the above model, crop yield has been considered as dependent variable (Y) while minimum temperature, maximum temperature and rainfall as exogenous variables (X).To this end, forecast of covariates using hybrid time domain approaches and neural network have been utilized in the fitted ARIMAX. Performance of developed forecast models have been thoroughly examined.

Hybrid linear time series approach using machine learning techniques

ARIMA model does not explain the nonlinearity component i.e. errors. Here, we can try to improve the performance of ARIMA model by explaining residuals through machine learning approaches like ANN and SVM. This method consists of two phases. In the first phase, the time series is analyzed by using ARIMA models. In the next phase, the residuals obtained in the previous phase are examined by ANN and then forecast values obtained from the ARIMA model are summed. Here, time delay neural network approach has been used to develop a new hybrid model to overcome the limitations of ARIMA in an attempt to yield more accurate results. A typical time delay neural network structure with one hidden layer is denoted by I:Hs:O*l*, where I is the number of nodes in input layer, s denotes the logistic sigmoid transfer function, O denotes number of nodes in the output layer and *I* indicates linear transfer function. Here, an ARIMA model is first used to model the linear patterns of time series. The residuals of the linear model will then contain only the nonlinear relationship. Therefore, in the second phase, the ANN and SVM are used to model the nonlinear patterns of ARIMA residuals. This hybrid approach is used to get better forecasts as compare to classical time series models. Artificial Neural Networks are flexible computing frameworks for modeling a broad range of nonlinear problems. One significant advantage of the ANN models over other classes of nonlinear models is that ANNs are universal approximators that can approximate a large class of functions with a high degree of accuracy. Their power comes from the parallel processing of the information from the data. No prior assumption of the model form is required in the model building process. The network model is largely determined by the characteristics of the data. Single hidden layer feed forward network is the most widely used model form for time series modeling and forecasting. The model is characterized by a network of three layers of simple processing units connected by a cyclic links. The relationship between the output and the inputs has the following mathematical representation:

$$y_t = w_0 + \sum_{j=1}^{Q} w_j g(w_{0j} + \sum_{i=1}^{P} w_{ij} y_{t-i}) + e_t$$

The logistic function is often used as the hidden layer activation function. Data normalization is often performed before the training process begins. When nonlinear transfer functions are used at the output nodes, the desired output values must be transformed to the range of the actual outputs of the network. Even if a linear output transfer function is used, it may still be advantageous to standardize the outputs as well as the inputs to avoid computational problems, to meet algorithm requirement and to facilitate network learning. In general data normalization is beneficial in terms of classification rate and mean squared errors, but the benefit diminishes as network and sample size increase. In addition data normalization usually slows down the training process. Normalization of the output values (targets) is usually independent of the normalization of the inputs. For time series modeling problems, however, the normalization of targets is typically performed together with the inputs. The choice of range to which inputs and targets are normalized depends largely on the activation function of output nodes, with typically [0, 1] for logistic function and [-1, 1] for hyperbolic tangent function. It should be noted that, as a result of normalizing the target values, the observed output of the network should be correspond to the normalized range. Thus, to interpret the results obtained from the network, the outputs must be rescaled to the original range. From the user's point of view, the accuracy obtained by ANNs should be based on the rescaled data sets. Performance measures is also be calculated on the rescaled outputs. Here, Zhang's hybrid approach (Zhang, 2003) has been employed. This approach considers time series (y_t) as a function of linear and nonlinear components. Hence

$$y_t = f(L_t, N_t)$$

Where L_t and N_t represents the linear and nonlinear component, respectively. As needs be the relationship between linear and nonlinear components, it can be written as following

$$y_t = L_t + N_t$$

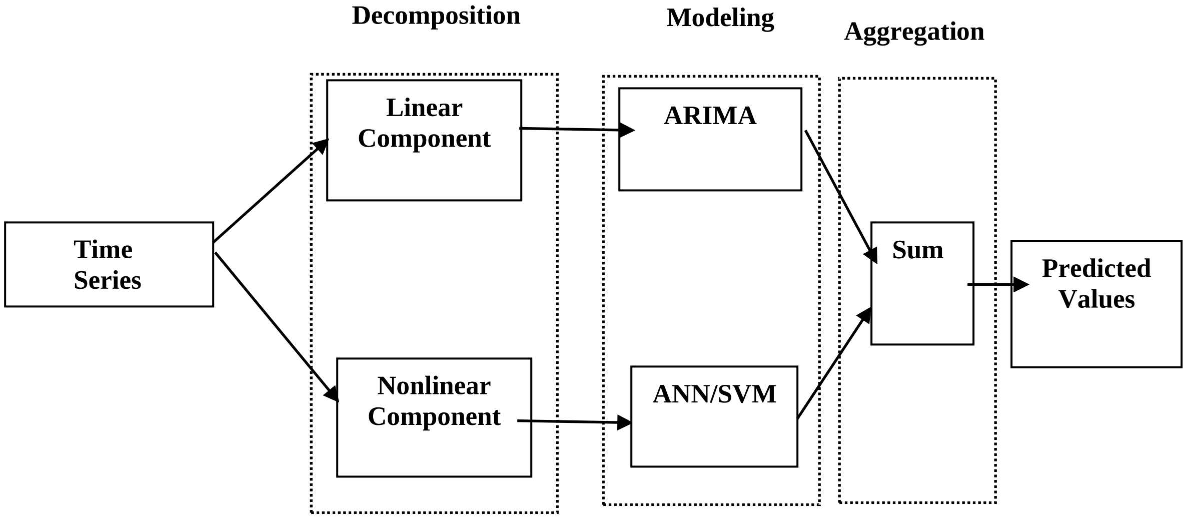
The main strategy of this approach is to model the linear and nonlinear components separately by different model. The methodology comprises of three steps. Initially, an ARIMA model is employed to fit the linear component. Let the prediction series provided by ARIMA model denoted as $\hat{L}_t$. In the second step, rather than predicting the linear component, the residuals denoted as e_t which are nonlinear in nature are predicted. The residuals can be gotten by subtracting the predicted value $\hat{L}_t$ from actual value of the considered time series y_t .

$$e_t = y_t - \hat{L}_t$$

Now the residuals are predicted employing an ANN model. Let the prediction series provided by ANN model denoted as $\hat{N}_t$. Eventually, the predicted linear and nonlinear components are combined to generate aggregate prediction.

$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

The ARIMA-ANN and ARIMA-SVM hybrid approach is graphically shown below



On the similar line we can compute ARIMAX-ANN and ARIMAX-SVM for improving the short term forecast of the time series data.

Out of sample forecast can be done through the identified neural network model based on minimum values of goodness of fit like MAPE. Using the forecasted out of sample rainfall values, we can get out of sample forecast of yield values using the selected ARIMAX model through forecast option of R-software.

Proposed approach for long term forecast

As we know univariate linear time series approaches like ARIMA or ARIMAX provides short term H-step ahead forecast. In H-step ahead forecasting, we learn H different models of the form

$$y_{t+h}=f_h(y_t,\ldots,y_{t-n+1})+\epsilon_{t+h},$$

For forecast h>H, we have proposed the following iterative steps for long term forecast through hybrid time series models through machine learning approaches:

1. Select the suitable ARIMA/ARIMAX model and obtain the fitted values of yield along with the residuals.

2. Test the residuals for nonlinearity, if residuals are nonlinear use machine learning techniques for modelling and forecasting of residuals.

3. Select the best ANN model for the residuals on the basis of minimum values of forecast accuracy measure (MAE or MAPE) and correct the fitted values of yield obtained via ARIMA/ARIMAX model through the fitted residuals estimated by the selected ANN model.

4. Compute the MAPE for the fitted values of yield through ARIMA/ARIMAX and hybrid approach.

5. If MAPE for hybrid approach is less than ARIMA/ARIMAX model, use the hybrid approach for long term forecast in the following way:

   i. Obtain the short term out of sample forecast of yield through the selected ARIMA model using the actual yield data.

   ii. Forecast the fitted residuals up to the desired forecast horizon by the suitable ANN model.

   iii. Obtain baseline data by correcting the short term forecast values of yield (obtained by ARIMA/ARIMA model) through the forecasted residuals using the selected ANN model.

   iv. Select suitable ARIMA model on the basis of baseline data and obtain short term forecast of the yield up to the desired forecast horizon.

   v. Consider the baseline data obtained as above for further long term forecast.

   vi. Repeat steps i-v until we get the forecast of the desired forecast horizon.

Satistical Modelling of Sensitive Issues on Successive Waves

Kumari Priyanka

Department of Mathematics, Shivaji College (University of Delhi),
New Delhi 110 027, India
email: *priyanka.ism@gmail.com*
Home Page: *https://sites.google.com/view/kumari-priyanka*

1 Introduction

Surveys dealing with sensitive questions, for example, drug usage, tax evasion, substance abuse, excessive gambling and AIDS pose particular problems. In such surveys, many respondents either refuse to participate or give false or evasive responses. Hence, in such situations the methods that protect anonymity are a solution. Such protection is built in to the two widely practiced ways. One is Randomized Response Technique(RRT) and other is Scrambled Response Technique(SRT). Whereas, recently a new technique called Item Sum Technique(IST) has been developed. Hence, here we intend to elaborate Item Sum Technique(IST) dealing with sensitive issues in successive sampling.

2 The item sum technique(IST)

The well known technique in sensitive characteristics estimation is the item count technique (ICT), however the ICT is generally applicable for qualitative variables only. Hence, the ICT was generalized by Chaudhuri & Christofides(2013) that can be used to estimate quantitative sensitive variable. Later Trappmann *et al.* (2014) named this generalized version of ICT as item sum technique (IST) and used it for estimating some quantitative sensitive variable. The algorithm for the IST is as follows:
From a random sample (say s), two random sub-samples (say s_{ll} and s_{sl}) are generated. The sub-sample s_{ll}, is confronted with a long list (LL) of items containing the sensitive question and a number of innocuous/non-sensitive questions.

However the respondents in sub-sample s_{sl} has been given a short list (SL) of items containing only the innocuous questions present in LL sample. The respondents in each samples are asked to report the total score of all the items given to them, without disclosing the individual scores for the items. The mean difference of the answers between the s_{ll} and s_{sl} is used as an unbiased estimator of the population mean of sensitive variable. It is to be noted that all sensitive and innocuous variables should be quantitative in nature and possibly measured on the same scale as that of the sensitive variable in the IST.

However, the decisive point in the IST is how to split the total sample in to the LL sample and SL sample. Trappmann *et al.* (2014) allocated the same number of units to each sample irrespective of the variation of items in the two list. However, Perri *et al.* (2018) advocated the requirement of optimum allocation of LL and SL samples. If the sensitive variable is also changing by time, which is often the scenario, then the IST may be modified to deal with sensitive issues on successive waves. For example if the sensitive variable is the amount spent on drugs such as cigarette, pan masala, etc. per month, by college students, then the non-sensitive variable may be taken as the total monthly pocket money received by them or the amount spent on purchasing books etc.. Similarly, if the sensitive variable is the number of abortion, then the non-sensitive variable may be the number of childrens or total number of members in that family etc.. The sensitive question together with non-sensitive questions will comprise of LL sample, however only non-sensitive questions will comprise of SL sample. There may be any number of non-sensitive question with a sensitive question to be used for LL sample and the same non-sensitive questions to be used for SL sample. But here we will consider one sensitive and one non-sensitive question case on successive waves.

3 Proposed IST Frame work in Successive Sampling Design

Consider a finite population P consisting of N identifiable units for sampling over two successive waves. Let x denote the quantitative sensitive variable at the first wave which changes to y at second wave. Similarly let t_1 be the non sensitive variable at the first wave which changes to t_2 at the second wave. Assume that x_i, y_i, t_{1i} and t_{2i} denotes the value of x, y, t_1 and t_2 respectively on the unit $i \epsilon\ P$. To estimate the population mean of quantitative sensitive variable $\bar{Y}$ at current wave using the IST, the sampling design is proposed as:
At first wave a sample of size n is drawn using simple random sample without replacement (SRSWOR) which has been split to s_{nll} and s_{nsl} samples called the LL-sample and SL-sample respectively. Now, at the second wave considering the partial overlap case, two independent samples are considered, one is a matched sample of size $m = n\lambda$ drawn as SRSWOR sub-sample from sample size n at first wave and second is a fresh sample of size $u = \ (n - m) \ = \ n\mu$, which is

drawn afresh at current wave. Further, the samples of sizes m and u are split in to corresponding LL-sample and SL-samples as s_{mll}, s_{msl}, s_{ull} and s_{usl} respectively. The response obtained from the respondents on two waves and the corresponding IST estimate based on different samples are presented in Table 1.

Table 1: Response received under IST

Wave	**Sample size**	**Response received**	**IST estimate**
I	n	$z_{1i}=\begin{cases} x_i+t_{1i} & if\ i\epsilon s_{nll} \\ t_{1i} & if\ i\epsilon s_{nsl} \end{cases}$	$\hat{\bar{x}}_n=\bar{z}_{1nll}-\bar{t}_{1nsl}$
II	m	$z_{1i}=\begin{cases} x_i+t_{1i} & if\ i\epsilon s_{mll} \\ t_{1i} & if\ i\epsilon s_{msl} \end{cases}$	$\hat{\bar{x}}_m=\bar{z}_{1mll}-\bar{t}_{1msl}$
	m	$z_{2i}=\begin{cases} y_i+t_{2i} & if\ i\epsilon s_{mll} \\ t_{2i} & if\ i\epsilon s_{msl} \end{cases}$	$\hat{\bar{y}}_m=\bar{z}_{2mll}-\bar{t}_{2msl}$
	u	$z_{2i}=\begin{cases} y_i+t_{2i} & if\ i\epsilon s_{ull} \\ t_{2i} & if\ i\epsilon s_{usl} \end{cases}$	$\hat{\bar{y}}_u=\bar{z}_{2ull}-\bar{t}_{2usl}$

Note: z_{ji} ; $j = 1, 2$ denote the observed response at first and second wave respectively on the i^{th} observation.

$\bar{z}_{jill}$; $j = 1, 2$; $i\ \epsilon \{n, m, u\}$ denote the mean of z_j in the long list (LL) samples.

$\bar{t}_{jisl}$; $j = 1, 2$; $i\ \epsilon\ \{n, m, u\}$ denote the mean of t_1 and t_2 in the short list (SL) samples.

4 IST Successive Difference Estimator

In order to utilize informations available from previous wave an IST difference type estimator $\mathbb{T}_{1m}$ is considered based on sample of size m retained from previous wave and the estimator based on sample of size u is the IST estimator $\mathbb{T}_u = \hat{\bar{y}}_u$. Combining the two estimators as the convex linear combinations, the final estimator for sensitive population mean at current wave is given by

$$\mathbb{T}_1 = \phi_1\mathbb{T}_u + (1-\phi_1)\mathbb{T}_{1m} \tag{1}$$

where $\mathbb{T}_u = \hat{\bar{y}}_u$ and $\mathbb{T}_{1m} = \hat{\bar{y}}_m + k(\hat{\bar{x}}_n - \hat{\bar{x}}_m)$; $\phi_1\ \epsilon\ [0, 1]$ and k is a scalar quantities to be chosen suitably.

5 IST Successive Regression Estimator

The another well known estimator in survey sampling theory is regression estimator. Hence, the estimator for the matched portion of sample have been chosen as regression type estimator given by $\mathbb{T}_{2m}$. The final estimator called

IST successive regression estimator for estimating sensitive population mean at current wave is given as

$$\mathbb{T}_2 \;=\; \phi_2 \mathbb{T}_u + (1-\phi_2)\mathbb{T}_{2m} \tag{2}$$

where $\mathbb{T}_{2m} = \hat{\bar{y}}_m + \hat{b}(m_{ll})(\hat{\bar{x}}_n - \hat{\bar{x}}_m)$ with $\hat{b}(m_{ll}) \;=\; \frac{s_{z_1 z_2}(m_{ll})}{s^2_{z_2}(m_{ll})}$ and $\phi_2 \;\epsilon\; [0,\;1]$ is a scalar quantity to be chosen suitably.

6 IST Successive General Class of Estimator

Many estimators such as ratio, product, exponential ratio etc., may be thought on similar lines for proposing estimator based on matched sample of size m. Therefore, in order to generalized the frame work, an IST general class of estimator has been proposed, so that the IST difference, IST regression and many others may be viewed as members of the proposed class of estimator. Hence, the final estimator in this case is given as

$$\mathbb{T}_3 \;=\; \phi_3 \mathbb{T}_u + (1-\phi_3)\mathbb{T}_{3m} \tag{3}$$

where, $\mathbb{T}_{3m} = g(\hat{\bar{y}}_m,\; \hat{\bar{x}}_m,\; \hat{\bar{x}}_n)$ is a function of $\hat{\bar{y}}_m,\; \hat{\bar{x}}_m$ and $\hat{\bar{x}}_n$. Following Priyanka & Trisandhya (2018), the function g is assumed such that it satisfies following conditions:

(i) The point $(\hat{\bar{y}}_m,\; \hat{\bar{x}}_m,\; \hat{\bar{x}}_n)$ assumes the value in a closed convex subset $\mathbb{R}^3$ of three dimensional real space containing the point $(\bar{Y},\; \bar{X},\; \bar{X})$.

(ii) The function $g\;(\hat{\bar{y}}_m,\; \hat{\bar{x}}_m,\; \hat{\bar{x}}_n)$ is continuous and bounded in $\mathbb{R}^3$.

(iii) $g(\bar{Y},\; \bar{X},\; \bar{X}) = \bar{Y}$ and $g_1(\bar{Y},\; \bar{X},\; \bar{X}) = \frac{\partial g(\hat{\bar{y}}_m,\; \hat{\bar{x}}_m,\; \hat{\bar{x}}_n)}{\partial \hat{\bar{y}}_m} = 1$.

(iv) The first and second order partial derivatives of $g\;(\hat{\bar{y}}_m,\; \hat{\bar{x}}_m,\; \hat{\bar{x}}_n)$ exist and are continuous and bounded in $\mathbb{R}^3$.

7 Analysis of IST estimators on Successive waves

To elucidate the performances of proposed IST estimators, the bias, variance/mean squared error of the proposed estimators $\mathbb{T}_i\;(i = 1,\; 2,\; 3)$ has been calculated as

$$\begin{aligned}
\mathbb{B}(\mathbb{T}_i) &= E\left(\mathbb{T}_i \;-\; \bar{Y}\right)\;;\; (i = 1,\; 2,\; 3) \\
&= E\left[\phi_i(\mathbb{T}_u - \bar{Y}) + (1-\phi_i)(\mathbb{T}_{im} - \bar{Y})\right] \\
&= \phi_i \mathbb{B}\left[\mathbb{T}_u\right] + (1-\phi_i)\mathbb{B}\left[\mathbb{T}_{im}\right]
\end{aligned}$$

Since, $\mathbb{T}_u$ is unbiased for $\bar{Y}$, so $\mathbb{B}(\mathbb{T}_u) = 0$. Therefore, the bias of estimator $\mathbb{T}_i$ is given as

$$\mathbb{B}(\mathbb{T}_i) = (1-\phi_i)\mathbb{B}(\mathbb{T}_{im}) \ ; \ (i = 1,\ 2,\ 3) \tag{4}$$

The variance of the estimator $\mathbb{T}_i$ is computed as

$$\begin{aligned}\mathbb{V}(\mathbb{T}_i) &= E\left(\mathbb{T}_i \ - \ \bar{Y}\right)^2 \ ; \ (i = 1,\ 2,\ 3)\\ &= E\left[\phi_i(\mathbb{T}_u - \bar{Y}) + (1-\phi_i)(\mathbb{T}_{im} - \bar{Y})\right]^2\\ &= \phi_i^2\mathbb{V}(\mathbb{T}_u) + (1-\phi_i)^2\mathbb{V}(\mathbb{T}_{im}) + 2\phi_i(1-\phi_i)cov(\mathbb{T}_u,\ \mathbb{T}_{im}) \end{aligned} \tag{5}$$

As $\mathbb{T}_u$ and $\mathbb{T}_{im}$ are based on two independent samples of sizes u and m respectively. So, $cov(\mathbb{T}_u,\ \mathbb{T}_{im}) = 0$. Therefore, the variance of estimator $\mathbb{T}_i$ becomes

$$\mathbb{V}(\mathbb{T}_i) = \phi_i^2\mathbb{V}(\mathbb{T}_u) + (1-\phi_i)^2\mathbb{V}(\mathbb{T}_{im}) \tag{6}$$

It can be seen that, $\mathbb{V}(\mathbb{T}_i)$ in equation (6) is a function of ϕ_i. So, it has been optimized with respect to ϕ_i and optimum value of ϕ_i is obtained as:

$$\phi_{iopt.} = \frac{\mathbb{V}(\mathbb{T}_{im})}{\mathbb{V}(\mathbb{T}_u) + \mathbb{V}(\mathbb{T}_{im})} \ ; \ i = 1,\ 2,\ 3 \tag{7}$$

Substituting the optimum value of ϕ_i from equation (7) in equation (6) the optimum variance of the proposed estimator $\mathbb{T}_i$ is computed as

$$\mathbb{V}(\mathbb{T}_i)_{opt.} = \frac{\mathbb{V}(\mathbb{T}_u) \times \mathbb{V}(\mathbb{T}_{im})}{\mathbb{V}(\mathbb{T}_u) + \mathbb{V}(\mathbb{T}_{im})} \ ; \ i \ = \ 1,\ 2,\ 3 \tag{8}$$

7.1 Bias and Variance of $\mathbb{T}_u$ and $\mathbb{T}_{im}$; $i \ = \ 1,\ 2,\ 3$

The estimator $\mathbb{T}_u$ is unbiased for $\bar{Y}$, hence, its variance is computed as

$$\mathbb{V}(\mathbb{T}_u) = \left(\frac{S_{z_2}^2}{u_{ll}}\right) + \left(\frac{S_{t_2}^2}{u_{sl}}\right) - \left(\frac{S_{z_2}^2 + S_{t_2}^2}{N}\right) \tag{9}$$

The estimator $\mathbb{T}_{1m}$ is also unbiased for $\bar{Y}$, so its variance is obtained as

$$\begin{aligned}\mathbb{V}(\mathbb{T}_{1m}) = &\left(\frac{1}{m_{ll}}\right)\left[k^2S_{z_1}^2 - 2k\rho_{z_1z_2}S_{z_1}S_{z_2} + S_{z_2}^2\right] + \left(\frac{1}{m_{sl}}\right)\left[k^2S_{t_1}^2 - 2kS_{t_1}S_{t_2}\rho_{t_1t_2} + S_{t_2}^2\right] +\\ &\left(\frac{1}{n_{ll}}\right)\left[2k\rho_{z_1z_2}S_{z_1}S_{z_2} - k^2S_{z_1}^2\right] + \left(\frac{1}{n_{sl}}\right)\left[2kS_{t_1}S_{t_2}\rho_{t_1t_2} - k^2S_{t_1}^2\right] -\\ &\left(\frac{1}{N}\right)\left[S_{z_2}^2 + S_{t_2}^2\right] \end{aligned} \tag{10}$$

The above expression for variance of $\mathbb{T}_{1m}$ contain an unknown constant k, hence, it is optimized with respect to k and optimum value of k is obtained as

$$k_{opt.} = \frac{(\frac{1}{m_{ll}} - \frac{1}{n_{ll}})S_{z_1}S_{z_2}\rho_{z_1z_2} + (\frac{1}{m_{sl}} - \frac{1}{n_{sl}})S_{t_1}S_{t_2}\rho_{t_1t_2}}{(\frac{1}{m_{ll}} - \frac{1}{n_{ll}})S^2_{z_1} + (\frac{1}{m_{sl}} - \frac{1}{n_{sl}})S^2_{t_1}}$$

Now, as the estimator $\mathbb{T}_{2m}$ and $\mathbb{T}_{3m}$ are biased for $\bar{Y}$, hence the expression for their bias and mean squared error have been computed under the following transformations: :
$\bar{z}_{2ull} = \bar{Z}_2\,(1+e_0)$, $\bar{z}_{2mll} = \bar{Z}_2\,(1+e_1)$, $\bar{z}_{1mll} = \bar{Z}_1\,(1+e_2)$, $\bar{z}_{1nll} = \bar{Z}_1\,(1+e_3)$, $\bar{t}_{2usl} = \bar{T}_2\,(1+e_4)$, $\bar{t}_{2msl} = \bar{T}_2\,(1+e_5)$, $\bar{t}_{1msl} = \bar{T}_1\,(1+e_6)$, $\bar{t}_{1nsl} = \bar{T}_1\,(1+e_7)$, $s^2_{z_2}(m_{ll}) = S^2_{z_2}\,(1+e_8)$, $s_{z_1z_2}(m_{ll}) = S_{z_1z_2}\,(1+e_9)$, $n_{rs} = \frac{\nu_{rs}}{(\nu_{20})^{r/2}(\nu_{02})^{s/2}}$, $\nu_{rs} = \frac{1}{n-1}\Sigma(z_{1i} - \bar{z}_1)^r(z_{2i} - \bar{z}_2)^s$.
such that, $E(e_i) = 0$; $|e_i| < 1$ where, $i = 0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8$ and 9.
Under the above transformations, retaining the terms up to first order of approximation, we have for bias and mean squared error of $\mathbb{T}_{2m}$ as

$$\mathbb{B}(\mathbb{T}_{2m}) = \left(\frac{1}{m_{ll}} - \frac{1}{n_{ll}}\right)\beta_{z_1z_2}\left[S_{z_1}n_{03} - S_{z_1}\frac{n_{12}}{\rho_{z_1z_2}}\right] \tag{11}$$

and

$$\mathbb{V}(\mathbb{T}_{2m}) = \left(\frac{1}{m_{ll}}\right)\left[S^2_{z_1}\beta^2_{z_1z_2} - 2S_{z_1}S_{z_2}\beta_{z_1z_2}\rho_{z_1z_2} + S^2_{z_2}\right] + \left(\frac{1}{m_{sl}}\right)\left[S^2_{t_2} + \beta^2_{z_1z_2}S^2_{t_1} - \right.$$
$$\left. 2\beta_{z_1z_2}S_{t_1}S_{t_2}\rho_{t_1t_2}\right] + \left(\frac{1}{n_{ll}}\right)\left[2S_{z_1}S_{z_2}\beta_{z_1z_2}\rho_{z_1z_2} - S^2_{z_1}\beta^2_{z_1z_2}\right] +$$
$$\left(\frac{1}{n_{sl}}\right)\left[2S_{t_1}S_{t_2}\rho_{t_1t_2}\beta_{z_1z_2} - S^2_{t_1}\beta^2_{z_1z_2}\right] - \left(\frac{1}{N}\right)\left[S^2_{z_2} + S^2_{t_2}\right] \tag{12}$$

Also, the bias and mean squared error of the estimator $\mathbb{T}_{3m}$ has been is derived under above considered transformations as:

$$\mathbb{T}_{3m} = g(\hat{\bar{y}}_m,\ \hat{\bar{x}}_m,\ \hat{\bar{x}}_n)$$

Expanding $g(\hat{\bar{y}}_m,\ \hat{\bar{x}}_m,\ \hat{\bar{x}}_n)$ about the point $K = (\bar{Y},\ \bar{X},\ \bar{X})$ using Taylor series expansion, retaining terms up to first order of approximations, we have

$$\mathbb{T}_{3m} = g\left[\bar{Y} + (\hat{\bar{y}}_m - \bar{Y}),\ \bar{X} + (\hat{\bar{x}}_m - \bar{X}),\ \bar{X} + (\hat{\bar{x}}_n - \bar{X})\right]$$
$$= (\hat{\bar{y}}_m + (\hat{\bar{x}}_m - \bar{X})G_2 + (\hat{\bar{x}}_n - \bar{X})G_3 + \left[(\hat{\bar{y}}_m - \bar{Y})^2G_{11} + (\hat{\bar{x}}_m - \bar{X})^2G_{22} + \right.$$
$$(\hat{\bar{x}}_n - \bar{X})^2G_{33} + (\hat{\bar{y}}_m - \bar{Y})(\hat{\bar{x}}_m - \bar{X})G_{12} + (\hat{\bar{y}}_m - \bar{Y})(\hat{\bar{x}}_n - \bar{X})G_{13} +$$
$$\left.\Big(\hat{\bar{x}}_m - \bar{X})(\hat{\bar{x}}_n - \bar{X})G_{23} + \ldots\right] \tag{13}$$

where,
$G_1 = \frac{\partial g}{\partial \hat{\bar{y}}_m}|_K = 1$, $G_2 = \frac{\partial g}{\partial \hat{\bar{x}}_m}|_K$, $G_3 = \frac{\partial g}{\partial \hat{\bar{x}}_n}|_K$, $G_{11} = \frac{1}{2}\frac{\partial^2 g}{\partial \hat{\bar{y}}_m^2}|_K = 0$, $G_{22} = \frac{1}{2}\frac{\partial^2 g}{\partial \hat{\bar{x}}_m^2}|_K$, $G_{33} = \frac{1}{2}\frac{\partial^2 g}{\partial \hat{\bar{x}}_n^2}|_K$, $G_{12} = \frac{1}{2}\frac{\partial^2 g}{\partial \hat{\bar{y}}_m \partial \hat{\bar{x}}_m}|_K$, $G_{13} = \frac{1}{2}\frac{\partial^2 g}{\partial \hat{\bar{y}}_m \partial \hat{\bar{x}}_n}|_K$ and $G_{23} = \frac{1}{2}\frac{\partial^2 g}{\partial \hat{\bar{x}}_m \partial \hat{\bar{x}}_n}|_K$.

Bias and Mean squared error of the class of estimator $\mathbb{T}_{3m}$ to the first order approximations are obtained as

$$\begin{aligned}\mathbb{B}(\mathbb{T}_{3m}) = {} & \frac{1}{m_{ll}}[S_{z_1}^2 G_{22} + \rho_{z_1z_2}S_{z_1}S_{z_2}G_{12}] + \frac{1}{m_{sl}}[S_{t_1}^2 G_{22} + \rho_{t_1t_2}S_{t_1}S_{t_2}G_{12}] + \\ & \frac{1}{n_{ll}}[S_{z_1}^2 G_{33} + \rho_{z_1z_2}S_{z_1}S_{z_2}G_{13} + S_{z_1}^2 G_{23}] + \frac{1}{n_{sl}}[S_{t_1}^2 G_{33} + \rho_{t_1t_2}S_{t_1}S_{t_2}G_{13} + \\ & S_{t_1}^2 G_{23}] - \frac{1}{N}[S_{t_1}^2(G_{22} + G_{33} + G_{23}) + S_{t_1}S_{t_2}\rho_{t_1t_2}(G_{12} + G_{13}) + \\ & S_{z_1}^2(G_{22} + G_{33} + G_{23}) + S_{z_1}S_{z_2}\rho_{z_1z_2}(G_{12} + G_{13})]\end{aligned} \tag{14}$$

and

$$\begin{aligned}\mathbb{V}(\mathbb{T}_{3m}) = {} & \left(\frac{1}{m_{ll}}\right)\left[S_{z_1}^2 G_2^2 + 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_2 + S_{z_2}^2\right] + \left(\frac{1}{m_{sl}}\right)\left[S_{t_1}^2 G_2^2 + \right. \\ & \left. 2S_{t_1}S_{t_2}\rho_{t_1t_2}G_2 + S_{t_2}^2\right] + \left(\frac{1}{n_{ll}}\right)\left[S_{z_1}^2 G_3^2 + 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_3 + 2G_2G_3S_{z_1}^2\right] + \\ & \left(\frac{1}{n_{sl}}\right)\left[S_{t_1}^2 G_3^2 + 2G_3S_{t_1}S_{t_2}\rho_{t_1t_2} + 2G_2G_3S_{t_1}^2\right] - \left(\frac{1}{N}\right)\left[S_{z_1}^2 G_2^2 + \right. \\ & 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_2 + S_{t_1}^2 G_2^2 + 2S_{t_1}S_{t_2}\rho_{t_1t_2}G_2 + S_{z_2}^2 + S_{t_2}^2 + S_{z_1}^2 G_3^2 + \\ & \left. 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_3 + 2G_2G_3S_{z_1}^2 + S_{t_1}^2 G_3^2 + 2S_{t_1}S_{t_2}\rho_{t_1t_2}G_3 + 2G_2G_3S_{t_1}^2\right]\end{aligned} \tag{15}$$

Clearly, we can see that equation (15) is a function of G_2 and G_3. So, after minimizing equation (15) by partially differentiating with respect to G_2 and G_3 respectively and equating to zero we get the optimized value of G_2 and G_3 as

$$G_{2opt.} = -\frac{\left[(\frac{1}{m_{ll}} - \frac{1}{n_{ll}})S_{z_1}S_{z_2}\rho_{z_1z_2} + (\frac{1}{m_{sl}} - \frac{1}{n_{sl}})S_{t_1}S_{t_2}\rho_{t_1t_2}\right]}{(\frac{1}{m_{ll}} - \frac{1}{n_{ll}})S_{z_1}^2 + (\frac{1}{m_{sl}} - \frac{1}{n_{sl}})S_{t_1}^2},$$

$$G_{3opt.} = -\frac{\left[(\frac{1}{n_{ll}} - \frac{1}{N})(S_{z_1}S_{z_2}\rho_{z_1z_2} + S_{z_1}^2 G_2) + (\frac{1}{n_{sl}} - \frac{1}{N})(S_{t_1}S_{t_2}\rho_{t_1t_2} + S_{t_1}^2 G_2)\right]}{(\frac{1}{n_{ll}} - \frac{1}{N})S_{z_1}^2 + (\frac{1}{n_{sl}} - \frac{1}{N})S_{t_1}^2}$$

7.2 Allocating LL & SL Sample using Trappmann *et al.* (2014) approach and Perri *et al.* (2018) approach

Since, in IST a sample is split in to LL sample and SL sample. Trappmann *et al.* (2014) considered equal number of units in both the samples irrespective of variability of the items in the two lists. Applying his approach on successive waves we have the following allocations:
$n_{ll} = n_{sl} = \frac{n}{2}$, $m_{ll} = m_{sl} = \frac{m}{2}$ and $u_{ll} = u_{sl} = \frac{u}{2}$

However, Perri *et al.* (2018) concluded that the estimates may be affected due to high variability of items in LL sample and SL sample. Hence, they proposed optimal sample size allocation to LL and SL samples by minimizing the variance of IST estimates under a budget constraints. Hence, modifying this ideas to work for allocating LL sample and SL sample on various samples at first and second wave assuming same budget allocation for each LL and SL samples we have: $n_{ll} = \frac{nS_{z_1}}{S_{z_1}+S_{t_1}} = n\beta_1$ (say), $n_{sl} = \frac{nS_{t_1}}{S_{z_1}+S_{t_1}} = n\beta_2$ (say), $u_{ll} = \frac{uS_{z_2}}{S_{z_2}+S_{t_2}} = u\beta_3$ (say), $u_{sl} = \frac{uS_{t_2}}{S_{z_2}+S_{t_2}} = u\beta_4$ (say), $m_{ll} = \frac{mS_{z_1}}{S_{z_1}+S_{t_1}} = m\beta_1$ (say) and $m_{sl} = \frac{mS_{t_1}}{S_{z_1}+S_{t_1}} = m\beta_2$ (say).

Using the assumptions of both the approaches the minimum variance of proposed estimators $\mathbb{T}_i$ has been obtained and are presented in Table 2.

Table 2:

Minimum Variance under Trappmann et al.(2014) approach	
$\mathbb{V}_t(\mathbb{T}_1)_{min.} = \left[\frac{\hat{\mu}_{1t}^2 J_{11} - \hat{\mu}_{1t} J_{12} + J_{13}}{\hat{\mu}_{1t}^2 J_{14} + \hat{\mu}_{1t} J_{15} + K_{11}}\right]$	with, $\hat{\mu}_{1t} = min\left\{\frac{-I_{12}+\sqrt{I_{12}^2+I_{11}I_{13}}}{I_{11}}, \frac{-I_{12}-\sqrt{I_{12}^2+I_{11}I_{13}}}{I_{11}}\right\} \epsilon\ [0,\ 1]$
$\mathbb{V}_t(\mathbb{T}_2)_{min.} = \left[\frac{\hat{\mu}_{2t}^2 J_{31} - \hat{\mu}_{2t} J_{32} + J_{33}}{\hat{\mu}_{2t}^2 J_{34} + \hat{\mu}_{2t} J_{35} + K_{31}}\right]$	with, $\hat{\mu}_{2t} = min\left\{\frac{-I_{32}+\sqrt{I_{32}^2+I_{31}I_{33}}}{I_{31}}, \frac{-I_{32}-\sqrt{I_{32}^2+I_{31}I_{33}}}{I_{31}}\right\} \epsilon\ [0,\ 1]$
$\mathbb{V}_t(\mathbb{T}_3)_{min.} = \left[\frac{\hat{\mu}_{3t}^2 J_{51} - \hat{\mu}_{3t} J_{52} + J_{53}}{\hat{\mu}_{3t}^2 J_{54} + \hat{\mu}_{3t} J_{55} + K_{51}}\right]$	with, $\hat{\mu}_{3t} = min\left\{\frac{-I_{52}+\sqrt{I_{52}^2+I_{51}I_{53}}}{I_{51}}, \frac{-I_{52}-\sqrt{I_{52}^2+I_{51}I_{53}}}{I_{51}}\right\} \epsilon\ [0,\ 1]$
Minimum Variance under Perri et al. (2018) approach	
$\mathbb{V}_p(\mathbb{T}_1)_{min.} = \left[\frac{\hat{\mu}_{1p}^2 J_{21} - \hat{\mu}_{1p} J_{22} + J_{23}}{\hat{\mu}_{1p}^2 J_{24} + \hat{\mu}_{1p} J_{25} + K_{21}}\right]$	with, $\hat{\mu}_{1p} = min\left\{\frac{-I_{22}+\sqrt{I_{22}^2+I_{21}I_{23}}}{I_{21}}, \frac{-I_{22}-\sqrt{I_{22}^2+I_{21}I_{23}}}{I_{21}}\right\} \epsilon\ [0,\ 1]$
$\mathbb{V}_p(\mathbb{T}_2)_{min.} = \left[\frac{\hat{\mu}_{2p}^2 J_{41} - \hat{\mu}_{2p} J_{42} + J_{43}}{\hat{\mu}_{2p}^2 J_{44} + \hat{\mu}_{2p} J_{45} + K_{41}}\right]$	with, $\hat{\mu}_{2p} = min\left\{\frac{-I_{42}+\sqrt{I_{42}^2+I_{41}I_{43}}}{I_{41}}, \frac{-I_{42}-\sqrt{I_{42}^2+I_{41}I_{43}}}{I_{41}}\right\} \epsilon\ [0,\ 1]$
$\mathbb{V}_p(\mathbb{T}_3)_{min.} = \left[\frac{\hat{\mu}_{3p}^2 J_{61} - \hat{\mu}_{3p} J_{62} + J_{63}}{\hat{\mu}_{3p}^2 J_{64} + \hat{\mu}_{3p} J_{65} + K_{61}}\right]$	with, $\hat{\mu}_{3p} = min\left\{\frac{-I_{62}+\sqrt{I_{62}^2+I_{61}I_{63}}}{I_{61}}, \frac{-I_{62}-\sqrt{I_{62}^2+I_{61}I_{63}}}{I_{61}}\right\} \epsilon\ [0,\ 1]$

where,

$$J_{j1} = K_{j0}K_{j3}f - K_{j0}K_{j4}f^2,\ J_{j2} = K_{j1}K_{j3} - f(K_{j0}K_{j4}f - K_{j0}K_{j2} - K_{j0}K_{j3}+$$

$$K_{j1}K_{j4}), J_{j3} = K_{j1}K_{j2} + K_{j1}K_{j3} - fK_{j1}K_{j4},\ J_{j4} = f(K_{j0} + K_{j4}) - K_{j3},$$

$$J_{j5} = K_{j2} + \ K_{j3} - K_{j1} - f(K_{j0} + K_{j4}),\ I_{j1} = J_{j1}J_{j5} + J_{j2}J_{j4},\ I_{j2} = J_{j1}K_{j1} -$$

$$J_{j4}J_{j3},\ \ I_{j3} = J_{j2}K_{j1} + \ J_{j3}J_{j5},\ \ K_{j0} = S_{z_2}^2 + S_{t_2}^2;\ \forall\ j\ =\ 1,\ 2,\ \ldots,\ 6\ ;$$

$$K_{j1} = 2S_{z_2}^2 + 2S_{t_2}^2\ ;\ \forall\ j = 1,\ 3,\ 5\ ;\ K_{j1} = \frac{S_{z_2}^2}{\beta_3} + \frac{S_{t_2}^2}{\beta_4}\ ;\ \forall\ j = 2,\ 4,\ 6\ ;$$

$$K_{12} = 2(k_t^2 S_{z_1}^2 - 2k_t S_{z_1} S_{z_2} \rho_{z_1 z_2} + S_{z_2}^2 + k_t^2 S_{t_1}^2 - 2k_t S_{t_1} S_{t_2} \rho_{t_1 t_2} + S_{t_2}^2),$$

$$K_{13} = 2(2k_t S_{z_1} S_{z_2} \rho_{z_1 z_2} - k_t^2 S_{z_1}^2 + 2k_t S_{t_1} S_{t_2} \rho_{t_1 t_2} - k_t^2 S_{t_1}^2),\ K_{14} = S_{z_2}^2 + S_{t_2}^2,$$

$$K_{22} = \frac{k_p^2 S_{z_1}^2 - 2k_p S_{z_1} S_{z_2} \rho_{z_1 z_2} + S_{z_2}^2}{\beta_1} + \frac{k_p^2 S_{t_1}^2 - 2k_p S_{t_1} S_{t_2} \rho_{t_1 t_2} + S_{t_2}^2}{\beta_2},$$

$$K_{23} = \frac{2k_p S_{z_1} S_{z_2} \rho_{z_1 z_2} - k_p^2 S_{z_1}^2}{\beta_1} + \frac{2k_p S_{t_1} S_{t_2} \rho_{t_1 t_2} - k_p^2 S_{t_1}^2}{\beta_2},\ K_{24} = S_{z_2}^2 + S_{t_2}^2$$

$$k_t\ =\ (S_{z_1} S_{z_2} \rho_{z_1 z_2} + S_{t_1} S_{t_2} \rho_{t_1 t_2}) \,/\, \left(S_{z_1}^2 + S_{t_1}^2\right),$$

$$k_p\ =\ \left(\frac{S_{z_1} S_{z_2} \rho_{z_1 z_2}}{\beta_1} + \frac{S_{t_1} S_{t_2} \rho_{t_1 t_2}}{\beta_2}\right) \Big/ \left(\frac{S_{z_1}^2}{\beta_1} + \frac{S_{t_1}^2}{\beta_2}\right),$$

$$K_{32} = 2(\frac{S_{z_1 z_2}^2}{S_{z_2}^4} S_{z_1}^2 - 2\frac{S_{z_1 z_2}}{S_{z_2}^2} S_{z_1} S_{z_2} \rho_{z_1 z_2} + \frac{S_{z_1 z_2}^2}{S_{z_2}^4} S_{t_1}^2 - 2\frac{S_{z_1 z_2}}{S_{z_2}^2} S_{t_1} S_{t_2} \rho_{t_1 t_2} + S_{z_2}^2 + S_{t_2}^2),$$

$$K_{33} = 2(2S_{z_1} S_{z_2} \rho_{z_1 z_2} \frac{S_{z_1 z_2}}{S_{z_2}^2} - S_{z_1}^2 \frac{S_{z_1 z_2}^2}{S_{z_2}^4} + 2S_{t_1} S_{t_2} \rho_{t_1 t_2} \frac{S_{z_1 z_2}}{S_{z_2}^2} - S_{t_1}^2 \frac{S_{z_1 z_2}^2}{S_{z_2}^4}),$$

$$K_{34} = S_{z_2}^2 + S_{t_2}^2,\ K_{42} = \left(S_{z_1}^2 \frac{S_{z_1 z_2}^2}{S_{z_2}^4} - 2S_{z_1} S_{z_2} \rho_{z_1 z_2} \frac{S_{z_1 z_2}}{S_{z_2}^2} + S_{z_2}^2\right) \frac{1}{\beta_1} +$$

$$\left(S_{t_1}^2 \frac{S_{z_1 z_2}^2}{S_{z_2}^4} - 2\frac{S_{z_1 z_2}}{S_{z_2}^2} S_{t_1} S_{t_2} \rho_{t_1 t_2} + S_{t_2}^2\right) \frac{1}{\beta_2},$$

$$K_{43} = \left(2S_{z_1} S_{z_2} \rho_{z_1 z_2} \frac{S_{z_1 z_2}}{S_{z_2}^2} - S_{z_1}^2 \frac{S_{z_1 z_2}^2}{S_{z_2}^4}\right) \frac{1}{\beta_1} + \left(2S_{t_1} S_{t_2} \rho_{t_1 t_2} \frac{S_{z_1 z_2}}{S_{z_2}^2} - S_{t_1}^2 \frac{S_{z_1 z_2}^2}{S_{z_2}^4}\right) \frac{1}{\beta_2},$$

$$K_{44} = S_{z_2}^2 + S_{t_2}^2,\ K_{52} = 2\left(S_{z_1}^2 G_{2t}^2 + 2\rho_{z_1 z_2} S_{z_1} S_{z_2} G_{2t} + S_{t_1}^2 G_{2t}^2 + 2S_{t_1} S_{t_2} \rho_{t_1 t_2} G_{2t} +\right.$$

$$\left. S_{z_2}^2 + S_{t_2}^2\right),\ K_{53} = 2\left[S_{z_1}^2 G_{3t}^2 + 2\rho_{z_1 z_2} S_{z_1} S_{z_2} G_{3t} + 2G_{2t} G_{3t} S_{z_1}^2 + S_{t_1}^2 G_{3t}^2 + 2G_{3t} S_{t_1} S_{t_2} \rho_{t_1 t_2} +\right.$$

$$\left. 2G_{2t} G_{3t} S_{t_1}^2\right],\ K_{54} = \left[S_{z_1}^2 G_{2t}^2 + 2\rho_{z_1 z_2} S_{z_1} S_{z_2} G_{2t} + S_{t_1}^2 G_{2t}^2 + 2S_{t_1} S_{t_2} \rho_{t_1 t_2} G_{2t} + S_{z_2}^2 + S_{t_2}^2 +\right.$$

$$\left. S_{z_1}^2 G_{3t}^2 + \ 2\rho_{z_1 z_2} S_{z_1} S_{z_2} G_{3t} + 2G_{2t} G_{3t} S_{z_1}^2 + S_{t_1}^2 G_{3t}^2 + \ 2G_{3t} S_{t_1} S_{t_2} \rho_{t_1 t_2} + 2G_{2t} G_{3t} S_{t_1}^2\right],$$

$$K_{62} = \frac{1}{\beta_1}\left(S_{z_1}^2 G_{2p}^2 + 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_{2p} + S_{z_2}^2\right) + \frac{1}{\beta_2}\left(S_{t_1}^2 G_{2p}^2 + 2S_{t_1}S_{t_2}\rho_{t_1t_2}G_{2p} + S_{t_2}^2\right),$$

$$K_{63} = \frac{1}{\beta_1}\left(S_{z_1}^2 G_{3p}^2 + 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_{3p} + 2G_{2p}G_{3p}S_{z_1}^2\right) + \frac{1}{\beta_2}\left(S_{t_1}^2 G_{3p}^2 + 2G_{3p}S_{t_1}S_{t_2}\rho_{t_1t_2} + \right.$$

$$\left. 2G_{2p}G_3S_{t_1}^2\right),$$

$$K_{64} = \left[S_{z_1}^2 G_{2p}^2 + 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_{2p} + S_{t_1}^2 G_{2p}^2 + 2S_{t_1}S_{t_2}\rho_{t_1t_2}G_{2p} + S_{z_2}^2 + S_{t_2}^2 + S_{z_1}^2 G_{3p}^2 + \right.$$

$$\left. 2\rho_{z_1z_2}S_{z_1}S_{z_2}G_{3p} + 2G_{2p}G_{3p}S_{z_1}^2 + S_{t_1}^2 G_{3p}^2 + 2G_{3p}S_{t_1}S_{t_2}\rho_{t_1t_2} + 2G_{2p}G_{3p}S_{t_1}^2\right],$$

$$G_{2t} = -(S_{z_1}S_{z_2}\rho_{z_1z_2} + S_{t_1}S_{t_2}\rho_{t_1t_2})/(S_{z_1}^2 + S_{t_1}^2),\ G_{3t} = -(S_{z_1}S_{z_2}\rho_{z_1z_2} + S_{z_1}^2 G_{2t} +$$

$$S_{t_1}S_{t_2}\rho_{t_1t_2} + S_{t_1}^2 G_{2t})/(S_{z_1}^2 + S_{t_1}^2),\ G_{2p} = -(\frac{S_{z_1}S_{z_2}\rho_{z_1z_2}}{\beta_1} + \frac{S_{t_1}S_{t_2}\rho_{t_1t_2}}{\beta_2})/(\frac{S_{z_1}^2}{\beta_1} + \frac{S_{t_1}^2}{\beta_2}),$$

$$G_{3p} = -(\frac{S_{z_1}S_{z_2}\rho_{z_1z_2} + S_{z_1}^2 G_{2p}}{\beta_1} + \frac{S_{t_1}S_{t_2}\rho_{t_1t_2} + S_{t_1}^2 G_{2p}}{\beta_2})/(\frac{S_{z_1}^2}{\beta_1} + \frac{S_{t_1}^2}{\beta_2})\ \ and\ f = \frac{n}{N}.$$

8 Efficiency Comparison

In order to compare various proposed IST estimators in successive sampling, the percent relative efficiencies have been computed for data considered in section (9) under Trappmann *et al.* (2014) as well as Perri *et al.* (2018) allocation designs as follows:

$$E_1 = \frac{\mathbb{V}_t(\mathbb{T}_1)_{min.}}{\mathbb{V}_p(\mathbb{T}_1)_{min.}} \times 100,\quad E_2 = \frac{\mathbb{V}_t(\mathbb{T}_2)_{min.}}{\mathbb{V}_p(\mathbb{T}_2)_{min.}} \times 100\ and\ E_3 = \frac{\mathbb{V}_t(\mathbb{T}_3)_{min.}}{\mathbb{V}_p(\mathbb{T}_3)_{min.}} \times 100.$$

9 Numerical Demonstration

Population Source:[Free access to data by Statistical Abstracts of United States]

To evaluate the performance of proposed IST successive sampling estimators, numerical illustrations has been supplemented using natural population. The population consists of $N = 51$ states. Let the aim be to estimate rate of abortion in year 2004. Therefore, for IST successive sampling frame work we consider:

y=Rate of abortions in the year 2004

x=Rate of abortions in the year 2000

t_1=Number of residents in the year 2000

t_2=Number of residents in the year 2004.

Clearly the rate of abortion is sensitive, however, rate of residents is non-sensitive. Hence, the data is suitable to be applied for IST frame work. Since same study variable "rate of abortion"has been observed for two different years 2000 and 2004, therefore, the considered data is suitable to be used for IST successive sampling frame work. The numerical calculations have been performed on above said data and results are represented in Table 3 :

Table 3: Emperical results

$\hat{\mu}_{1p}$	$\hat{\mu}_{1t}$	$\hat{\mu}_{2p}$	$\hat{\mu}_{2t}$	$\hat{\mu}_{3p}$	$\hat{\mu}_{3t}$	E_1	E_2	E_3
0.8424	0.8359	0.7617	0.7626	*	*	114.9334	115.8558	−

Note: $'*'$ indicates that the optimum value of fraction of sample to be drawn afresh do not exist and $'-'$ denote corresponding percent relative efficiency cannot be computed.

The value of E_i $(i = 1, 2)$ are observed to be more than 100, this indicates that optimum allocation design by Perri *et al.* (2018) is preferable over Trappmann *et al.* (2014) design. Therefore, the further numerical analysis has been carried out using Perri et al. (2018) allocation design.

10 Simulation Study

An extensive simulation study has been carried out using Monte Carlo simulation for the data mentioned in section (9). The $5,000$ different Monte carlo replications have been observed. The process is also repeated for different combination of constants termed as sets. The variance/mean squared error of the proposed estimators $\mathbb{T}_1,\ \mathbb{T}_2$ and $\mathbb{T}_3$ has been computed under Perri *et al.* (2018) allocation design and are denoted by $\mathbb{V}_p(\mathbb{T}_1),\ \mathbb{V}_p(\mathbb{T}_2)$ and $\mathbb{V}_p(\mathbb{T}_3)$ respectively. The percent relative efficiencies of IST successive difference and regression estimators with respect to IST successive general class of estimator have been computed as:

$$E_{s1} = \frac{\mathbb{V}_p(\mathbb{T}_1)}{\mathbb{V}_p(\mathbb{T}_3)} \times 100 \ \ and\ E_{s2} = \frac{\mathbb{V}_p(\mathbb{T}_2)}{\mathbb{V}_p(\mathbb{T}_3)} \times 100$$

The procedure have been repeated for three different combinations of constants, termed as different sets given as:
I: $n = 24,\ u = 04, m = 20$
II: $n = 24,\ u = 08, m = 16$
III: $n = 24,\ u = 10, m = 14$.
The outcomes of simulation results are summarized in Figure 10.1 and Figure 10.2 respectively.

11 Direct Method

It is to be noted that under IST the estimators are less efficient than estimators obtained using direct questioning. Hence, in order to identify the amount of loss

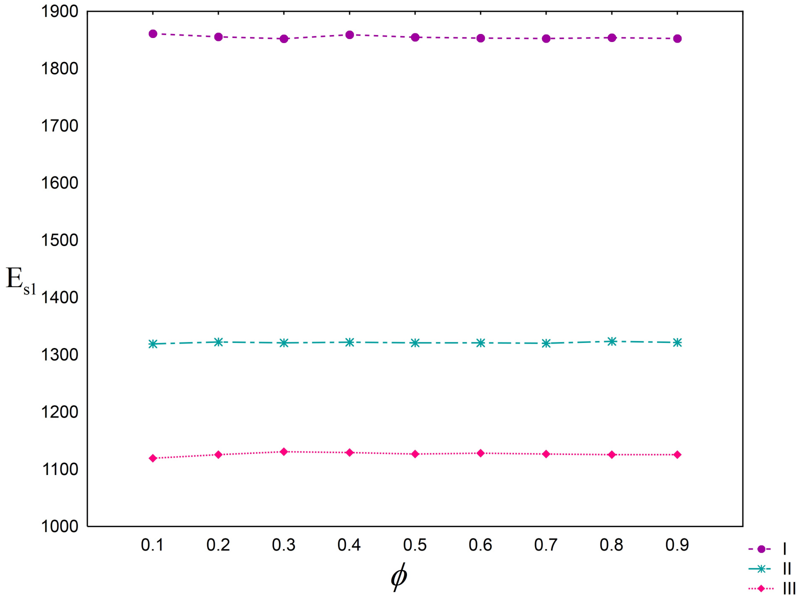


Figure 10.1: Simulated Percent relative efficiency of the IST general class of estimator with respect to IST difference estimator for three different sets

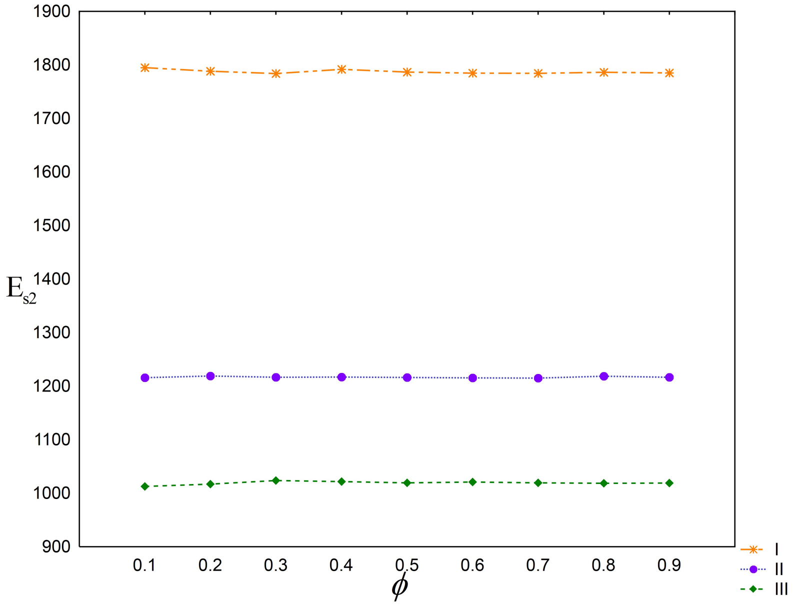


Figure 10.2: Simulated Percent relative efficiency of the IST general class of estimator with respect to IST regression estimator for three different sets

we compare the class of estimator $\mathbb{T}_3$ with respect to corresponding direct method. The estimator under direct questioning method is given as

$$\mathbb{T}_D = \chi \mathbb{T}_{uD} \ + \ (1-\chi)\mathbb{T}_{3mD} \ ; \ \chi \ \epsilon \ [0, \ 1] \tag{16}$$

where

$$\mathbb{T}_{uD} = \ \bar{y}_u \tag{17}$$

$$\mathbb{T}_{3mD} = \ d(\bar{y}_m, \ \bar{x}_m, \ \bar{x}_n) \tag{18}$$

where, $\mathbb{T}_{3mD}$ follow similar regularity conditions as stated in section (6). The, minimum mean squared error of the class of estimator $\mathbb{T}_D$ is obtained as

$$\mathbb{V}(\mathbb{T}_D)_{min.} = \left[\frac{-\hat{\mu}_D^2 J_{d1} + \hat{\mu}_D J_{d2} + J_{d3}}{\hat{\mu}_D^2 J_{d4} - \hat{\mu}_D J_{d5} + K_{d1}}\right] \tag{19}$$

with,

$$\hat{\mu}_D = min\left\{\frac{I_{d2} + \sqrt{I_{d2}^2 - I_{d1}I_{d3}}}{I_{d1}}, \ \frac{I_{d2} - \sqrt{I_{d2}^2 - I_{d1}I_{d3}}}{I_{d1}}\right\} such \ that \ \hat{\mu}_D \ \epsilon \ [0, \ 1] \tag{20}$$

where,
$J_{d1} = K_{d1}K_{d3}f^2$, $J_{d2} = f(K_{d1}K_{d3} - K_{d1}K_{d2} + fK_{d1}K_{d3})$, $J_{d3} = K_{d1}K_{d2} - fK_{d1}K_{d3}$, $J_{d4} = f(K_{d1} + K_{d3})$, $J_{d5} = K_{d1} - K_{d2} + f(K_{d1} + K_{d3})$, $I_{d1} = J_{d1}J_{d5} - J_{d2}J_{d4}$, $I_{d2} = J_{d3}J_{d4} + J_{d1}K_{d1}$, $I_{d3} = J_{d2}K_{d1} + J_{d3}J_{d5}$, $K_{d1} = S_y^2$, $K_{d2} = S_y^2 + S_xS_yD_2^2 + 2S_y^2D_2\rho_{yx}$, $K_{d3} = S_y^2 + S_x^2D_2^2 + 2S_yS_x\rho_{yx}D_2$ and $D_2 = (-S_yS_x\rho_{yx})/(S_x^2)$.

Further the simulated ratio of the mean squared error of $\mathbb{T}_D$ and $\mathbb{T}_3$ have been computed by considering $5,000$ different samples using Monte carlo simulation study for different sets and results are presented in Figure 11.1 and Figure.11.2 respectively

$$Ratio \ = \frac{\mathbb{V}(\mathbb{T}_D)}{\mathbb{V}_p(\mathbb{T}_3)} \tag{21}$$

12 Results and Discussions

Following interpretations can be drawn from empirical and simulation results:

1. It has been observed that IST is feasible in successive sampling to handle

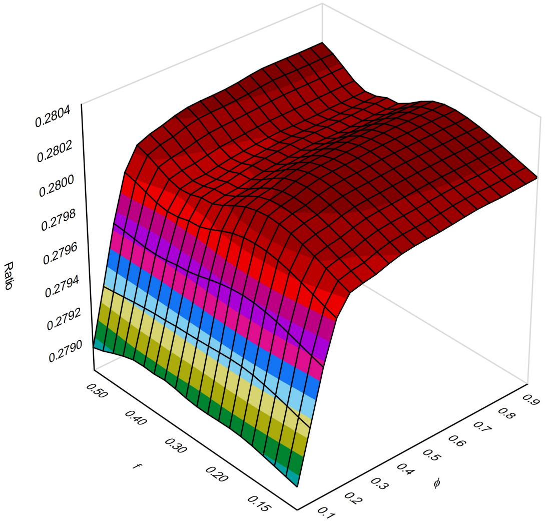


Figure 11.1: Ratio of mean squared error of $\mathbb{T}_3$ (under optimum allocation design) with respect to direct method under IST in two wave successive sampling for Set-I

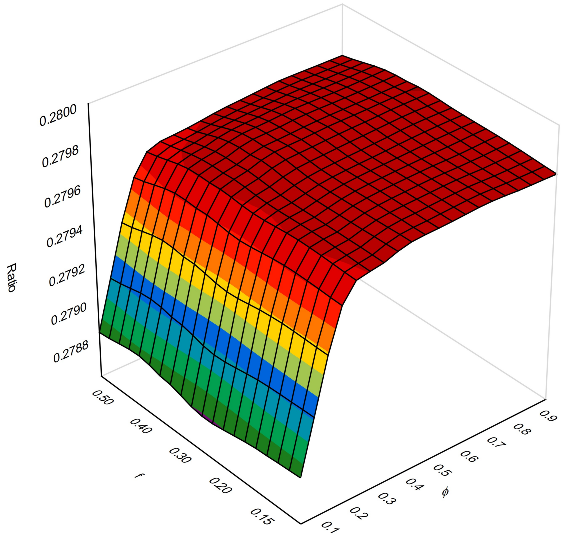


Figure 11.2: Ratio of mean squared error of $\mathbb{T}_3$ (under optimum allocation design) with respect to direct method under IST in two wave successive sampling for Set-II

sensitive issues on successive waves.

2. From Table 3, it is clear that E_1 and E_2 both are coming out to be greater than 100, this implies that optimum allocation design by Perri *et al.* (2018) is more efficient than allocation by Trappmann *et al.* (2014) design in two wave successive sampling. It is to be noted that for the considered data the optimum fraction to be drawn afresh do not exist for IST successive general class of estimator, so corresponding efficiency cannot be computed. Hence, in order to check the validity of IST successive general class of estimator simulation has been carried out with several choices of parameters.

3. Simulation results in Figure 10.1 and Figure 10.2 , justifies that E_{s_1} and E_{s_2} are greater than 100 for all three considered sets. This indicates that IST successive general class of estimators is more efficient than IST successive regression and IST successive difference estimators. However, $E_{s_1} > E_{s_2}$ indicate that IST successive difference estimator is better than IST successive regression estimator. Also, as ϕ increases, the simulated percent relative efficiency increases, this is in accordance with the theory of successive sampling.

4. From Figure 11.1 and Figure 11.2, it is observed that as ϕ increases simulated ratio of mean squared error of direct method and IST successive general class of estimator increases. The values indicate loss in precision of IST general class of estimator over direct method. Since the issues under consideration are sensitive so direct method will not serve the purpose as privacy of respondents need to be addressed. Hence, despite of loss in precision IST need to be preferred over direct method for sensitive issues on successive waves.

The IST on successive waves enables us to estimate the population mean of stigmatized quantitative variable using innocuous informations, there by reducing social desirability response bias and providing privacy to some extent. Out of the three proposed IST successive estimators, the IST successive general class of estimator under both Trappmann *et al.* (2014) allocation design as well as Perri *et al.* (2018) allocation design have been proved to be more efficient than other two. The optimum allocation of LL and SL samples by Perri *et al.* (2018) have been found to be more fruitful in successive sampling than allocation due to Trappmann *et al.* (2014). While comparing with direct method, certain amount of loss in precision is observed but that is realistic as the survey issues are sensitive so there may be chances of complete refusal or partial refusal if we apply direct method. However, using IST on successive wave atleast estimation of sensitive issues are possible. Therefore it can be concluded that the proposed IST successive estimators not only provide comfort and satisfaction to the respondents in terms of privacy protection but will also be a methodological advancement in literature related to successive sampling dealing sensitive issues.

13 Extension of IST set up to General Sampling Design on Successive Waves

Let us consider a finite population $U = (U_1,\, U_2,\, \ldots,\, U_N)$ consisting of N different identifiable units, which has been sampled over two successive waves. Let $y_1(y_2)$ denote the sensitive variable at first(second) wave. Aim is to estimate population mean of sensitive variable at current wave in an IST setting in successive sampling. In doing so, a sample of size n is drawn at first wave with design d_1 having probability $P_{d_1}(s_n)$, which is further split in to two independent samples say s_{nll} and s_{nsl} called the long list (LL) sample and short list (SL) sample respectively with design d_{1ll} & d_{1sl} and corresponding probabilities $P_{d_{1ll}}(s_{nll})$ & $P_{d_{1sl}}(s_{nsl})$ respectively. Now, considering the partial overlap case on second wave, two independent samples are drawn, one is matched sample of size m drawn as a sub-sample from sample s_{nll} with design d_2 having probability $P_{d_2}(s_m|s_{nll})$. The sample s_m is further split in to independent sub-samples s_{mll} and s_{msl} with design d_{2ll} & d_{2sl} with corresponding probabilities $P_{d_{2ll}}(s_{mll}|s_{nll})$ & $P_{d_{2sl}}(s_{msl}|s_{nll})$ respectively. However, other sample of size u is drawn afresh at current wave with design d_3 with probability $P_{d_3}(s_u|s_n^c)$. For IST setting the sample s_u is further split in to two independent sub-samples s_{ull} and s_{usl} with design d_{3ll} & d_{3sl} having probability $P_{d_{3ll}}(s_{ull})$ & $P_{d_{3sl}}(s_{usl})$ respectively. The first and second order positive inclusion probabilities for different samples are shown in Table 4.

Table 4: Inclusion Probabilities

		First order inclusion probability	Second order inclusion probability
n	n_{ll}	$\pi'_{ill} = \sum\limits_{s_{nll}\ni i} P_{d_{1ll}}(s_{nll})$	$\pi'_{ijll} = \sum\limits_{s_{nll}\ni i\&j} P_{d_{1ll}}(s_{nll})$
	n_{sl}	$\pi'_{isl} = \sum\limits_{s_{nsl}\ni i} P_{d_{1sl}}(s_{nsl})$	$\pi'_{ijsl} = \sum\limits_{s_{nsl}\ni i\&j} P_{d_{1sl}}(s_{nsl})$
m	m_{ll}	$\pi_{ill}\vert s_{nll} = \sum\limits_{s_{mll}\ni i} P_{d_{2ll}}(s_{mll})$	$\pi_{ijll}\vert s_{nll} = \sum\limits_{s_{mll}\ni i\&j} P_{d_{2ll}}(s_{mll})$
	m_{sl}	$\pi_{isl}\vert s_{nll} = \sum\limits_{s_{msl}\ni i} P_{d_{2sl}}(s_{msl})$	$\pi_{ijsl}\vert s_{nll} = \sum\limits_{s_{msl}\ni i\&j} P_{d_{2sl}}(s_{msl})$
u	u_{ll}	$\pi_{ill} = \sum\limits_{s_{ull}\ni i} P_{d_{3ll}}(s_{ull})$	$\pi_{ijll} = \sum\limits_{s_{ull}\ni i\&j} P_{d_{3ll}}(s_{ull})$
	u_{sl}	$\pi_{isl} = \sum\limits_{s_{usl}\ni i} P_{d_{3sl}}(s_{usl})$	$\pi_{ijsl} = \sum\limits_{s_{usl}\ni i\&j} P_{d_{3sl}}(s_{usl})$

Note: $s_k \ni i$ denotes that the sum is over those samples s_k that contain the given i, where, $k \in \{n_{ll},\, n_{sl},\, m_{ll},\, m_{sl},\, u_{ll},\, u_{sl}\}$.

Now, applying IST setting, all LL samples are confronted with two questions, out of which one refers to non-sensitive variable and other one is sensitive variable under study. However, all SL samples receive only one innocuous question which was used in LL sample. It is assumed that sensitive and non-sensitive items are quantitative in nature. Respondents in each sample are requested to report the total score of items without revealing the individual scores for the items.

Let $t_1(t_2)$ denote the non-sensitive item at first(second) wave. The response obtained from the i^{th} respondent at different waves in relevant samples in IST set up are presented in Table 5.

Table 5: Response received

Wave	**Sample size**	**Response received**
I	n	$z_{1i} = \begin{cases} y_{1i} + t_{1i} & if\ i \in s_{nll} \\ t_{1i} & if\ i \in s_{nsl} \end{cases}$
	m	$z_{1i} = \begin{cases} y_{1i} + t_{1i} & \text{if } i \in s_{mll} \\ \text{t}_{1i} & \text{if } i \in s_{msl} \end{cases}$
II	m	$z_{2i} = \begin{cases} y_{2i} + t_{2i} & \text{if } i \in s_{mll} \\ \text{t}_{2i} & \text{if } i \in s_{msl} \end{cases}$
	u	$z_{2i} = \begin{cases} y_{2i} + t_{2i} & \text{if } i \in s_{ull} \\ \text{t}_{2i} & \text{if } i \in s_{usl} \end{cases}$

Note: z_{ji} ; $j = 1, 2$ denote the observed response at first and second wave respectively on the i^{th} observation.

The Sampling designs basic weights for the sample of size n, m and u are described in Table 6.

Table 6: Sampling Design

Sample	**Sample Size**	Sampling design basic weight for selecting i^{th} unit in corresponding sample	
		LL-Sample	SL-Sample
s_n	n	$a_{1i}(ll) = \frac{1}{\pi'_{ill}}$	$a_{1i}(sl) = \frac{1}{\pi'_{isl}}$
s_m	m	$a_i^*(ll) = \frac{1}{\pi'_{ill}\pi_{ill\|s_{nll}}}$	$a_i^*(sl) = \frac{1}{\pi'_{isl}\pi_{isl\|s_{nll}}}$
s_u	u	$b_i^*(ll) = \frac{1}{(\pi'_{ill})^c\pi_{ill\|(s_n)^c}}$	$b_i^*(sl) = \frac{1}{(\pi'_{isl})^c\pi_{isl\|(s_n)^c}}$

Based on above sampling design basic weight, we intend to modify Horvitz-Thomson(1952) estimator in IST successive sampling setup to work for estimation of sensitive population mean under generic sampling design on successive waves.

14 IST successive HT-type Estimator

In IST successive sampling setup, to estimate sensitive population mean at current wave in two wave successive sampling, in general two samples are available.
Based on fresh sample of size u, the IST successsive Horvitz-Thomson type estimator is proposed as

$$T_{Hu} = \frac{1}{N}\sum_{i\epsilon s_{ull}} b_i^*(ll)z_{2i} - \frac{1}{N}\sum_{i\epsilon s_{usl}} b_i^*(sl)t_{2i} \tag{22}$$

Similarly, based on matched sample of size m, the IST successive HT-type estimator is proposed as

$$T_{Hm} = \frac{1}{N}\sum_{i\epsilon s_{mll}} a_i^*(ll)z_{2i} - \frac{1}{N}\sum_{i\epsilon s_{msl}} a_i^*(sl)t_{2i} \tag{23}$$

Now, Considering the convex linear combination of proposed estimators T_{Hu} and T_{Hm} from (22) and (23) respectively, we have the final IST successive Horvitz-Thomson type estimator T_H of sensitive population mean $\bar{Y}_2$ as

$$T_H = \phi_H T_{Hu} + (1-\phi_H)T_{Hm},\ where\ \phi_H \in [0,\ 1] \tag{24}$$

14.1 Remark. Selection of ϕ_H is purely dependent on usage of estimates. If fresh estimate of population mean is required on each wave then the estimator T_{Hu} is suitable. They may be utilized by choosing ϕ_H as 1 (or close to 1). While for reliable estimates of change in population mean from one wave to another, selecting ϕ_H as 0 (or close to 0) shows more emphasis to estimator T_{Hm}. Hence, a suitable choice of ϕ_H is desired for affirming both the situations at the same time.

14.2 Remark. It is to be noted that, in order to propose IST successive Horvitz-Thomson type estimator only informations at current occasion have been utilized. However, there is a scope to use information from previous wave as an auxiliary information to be used at current wave together with the availability of additional auxiliary variables. To avoid the impact in inference, which may creep due to bad sample selection and also to make use of additional auxiliary information, calibration approach is more suitable. Hence, in next section we propose IST successive calibration estimator to estimate sensitive population mean at current wave in two waves successive sampling.

15 IST Successive Calibration Estimator

Let $\mathbb{x}_i = (x_1,\ x_2, \ldots,\ x_p)^t$ be the p-additional auxiliary variables available at both the waves. Utilizing these p-additional auxiliary variables, a new IST calibration estimator T_{cu} and T_{cm} have been proposed which are based on sample of sizes u and m respectively for the estimation of sensitive population mean at current(second) wave in two waves successive sampling.

15.1 Estimator based on fresh sample of size u

The proposed IST calibration estimator T_{cu} using p-additional auxiliary variables based on sample of size u drawn afresh at current wave is described by replacing the basic design weights $b_i^*(ll)$ & $b_i^*(sl)$ for LL & SL samples by the new weights w_{ulli} & w_{usli} respectively. Therefore, the proposed IST calibration estimator for the estimation of sensitive population mean based on fresh sample becomes

$$T_{cu} = \frac{1}{N}\sum_{i\in s_{ull}} w_{ulli} z_{2i} - \frac{1}{N}\sum_{i\in s_{usl}} w_{usli} t_{2i} \tag{25}$$

To obtain the calibrated weight w_{ulli} of LL-sample, we minimize the chi-square type function

$$\sum_{i\in s_{ull}} \frac{\left(w_{ulli} - b_i^*(ll)\right)^2}{q_{ulli} b_i^*(ll)} \tag{26}$$

subject to calibration constraints

$$\frac{1}{N}\sum_{i\in s_{ull}} w_{ulli}\mathbb{x}_i = \bar{\mathbb{X}} \tag{27}$$

with q_{ulli} being known positive constant unrelated to $b_i^*(ll)$ and $\mathbb{x}_i = (x_{1i},\ x_{2i},\ \ldots,\ x_{pi})^t$ and $\bar{\mathbb{X}} = (\bar{X}_1,\ \bar{X}_2,\ \ldots,\ \bar{X}_p)^t$.
Now, Minimizing chi-square type function subject to constraints given in equation (27) lead to the calibrated weight given by

$$w_{ulli} = b_i^*(ll) + b_i^*(ll) q_{ulli} N \left[\frac{\left(\bar{\mathbb{X}} - \frac{1}{N}\sum\limits_{i\in s_{ull}} b_i^*(ll)\mathbb{x}_i\right)\mathbb{x}_i}{\sum\limits_{i\in s_{ull}} \mathbb{x}_i\mathbb{x}_i^t q_{ulli} b_i^*(ll)}\right] \tag{28}$$

Similarly, from SL sample the calibrated weight w_{usli} is obtained as

$$w_{usli} = b_i^*(sl) + b_i^*(sl) q_{usli} N \left[\frac{\left(\bar{\mathbb{X}} - \frac{1}{N}\sum\limits_{i\in s_{usl}} b_i^*(sl)\mathbb{x}_i\right)\mathbb{x}_i}{\sum\limits_{i\in s_{usl}} \mathbb{x}_i\mathbb{x}_i^t q_{usli} b_i^*(sl)}\right] \tag{29}$$

Therefore, after substituting the obtained calibrated weights w_{ulli} and w_{usli} in equation(25) we have the final IST calibrated estimator T_{cu} as

$$T_{cu} = \left[\frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)z_{2i} + \hat{\mathbb{B}}_{ull}\left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)\mathbb{x}_i\right)^t\right] - \left[\frac{1}{N}\sum_{i\in s_{usl}} b_i^*(sl)t_{2i} + \right.$$

$$\left.\hat{\mathbb{B}}_{usl}\left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{usl}} b_i^*(sl)\mathbb{x}_i\right)^t\right] \tag{30}$$

where

$$\hat{\mathbb{B}}_{ull} = \left(\sum_{i\in s_{ull}} b_i^*(ll)q_{ulli}\mathbb{x}_i\mathbb{x}_i^t\right)^{-1}\left(\sum_{i\in s_{ull}} b_i^*(ll)q_{ulli}\mathbb{x}_i z_{2i}\right)$$

and

$$\hat{\mathbb{B}}_{usl} = \left(\sum_{i\in s_{usl}} b_i^*(sl)q_{usli}\mathbb{x}_i\mathbb{x}_i^t\right)^{-1}\left(\sum_{i\in s_{usl}} b_i^*(sl)q_{usli}\mathbb{x}_i t_{2i}\right)$$

and q_{usli} being known positive constant unrelated to $b_i^*(sl)$ and $\bar{\mathbb{X}}$. The estimator T_{cu} in equation (30) can be written as:

$$T_{cu} = T_{cull} - T_{cusl} \tag{31}$$

15.2 Estimator based on matched sample

To improve the performance of the estimators on the current wave, it is well-known practice to utilize the information gathered on the previous occasion as auxiliary information in addition to p-additional auxiliary variables. Based on sample of size m at current(second) wave with the new calibrated weight, the IST calibration estimator is proposed as

$$T_{cm} = \frac{1}{N}\sum_{i\in s_{mll}} w_{mlli}z_{2i} - \frac{1}{N}\sum_{i\in s_{msl}} w_{msli}t_{2i} \tag{32}$$

Now, to obtain the calibrated weight w_{mlli} of LL sample we minimize the chi-square function

$$\sum_{i\in s_{mll}} \frac{(w_{mlli} - a_i^*(ll))^2}{q_{mlli}a_i^*(ll)} \tag{33}$$

subject to calibration constraints

$$\frac{1}{N}\sum_{i\in s_{mll}} w_{mlli}y_{1i} \;=\; \bar{y}^{c}_{1nll}, \tag{34}$$

and

$$\frac{1}{N}\sum_{i\in s_{mll}} w_{mlli}\mathbb{x}_i = \bar{\mathbb{X}} \tag{35}$$

with q_{mlli} as known positive constant unrelated to $a_i^*(ll)$ and $\bar{\mathbb{X}}$. Following similar procedure, we obtain the IST Calibration estimator $\bar{y}^c_{1nll}$ based on sample of size n drawn at first wave and used as auxiliary information at current(second) wave as.

$$\bar{y}^{c}_{1nll} = \left[\frac{1}{N}\sum_{i\in s_{nll}} a_{1i}(ll)z_{1i} + \hat{\mathbb{B}}_{nll}\left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{nll}} a_{1i}(ll)\mathbb{x}_i\right)^{t}\right] \tag{36}$$

with

$$\hat{\mathbb{B}}_{nll} = \left(\sum_{i\in s_{nll}} a_{1i}(ll)q_{nlli}\mathbb{x}_i\mathbb{x}_i^t\right)^{-1}\left(\sum_{i\in s_{nll}} a_{1i}(ll)q_{nlli}\mathbb{x}_i z_{1i}\right)$$

where q_{nlli} is known positive constant unrelated to $a_{1i}(ll)$ and $\bar{\mathbb{X}}$.
Now, Minimizing chi-square function in equation (33) subject to constraints in equation (34) and equation (35) lead to the calibrated weight given by

$$w_{mlli} = a_i^*(ll) + a_i^*(ll)q_{mlli}N\left[\frac{\left(\bar{\mathbb{X}}_{lm} - \frac{1}{N}\sum\limits_{i\in s_{mll}} a_i^*(ll)\mathbb{x}_{mi}\right)\mathbb{x}_{mi}}{\sum\limits_{i\in s_{mll}} \mathbb{x}_{mi}\mathbb{x}^t_{mi}q_{mlli}a_i^*(ll)}\right] \tag{37}$$

Where, $\mathbb{x}_{mi} = (y_{1i},\ x_{1i},\ x_{2i},\ \ldots,\ x_{pi})^t$, $\bar{\mathbb{X}}_{lm} = (\bar{y}^c_{1nll},\ \bar{X}_1,\ \bar{X}_2,\ \ldots,\ \bar{X}_p)^t$.
Now, by following similar procedure for SL-sample, the calibrated weight w_{msli} is given as

$$w_{msli} = a_i^*(sl) + a_i^*(sl)q_{msli}N\left[\frac{\left(\bar{\mathbb{X}}_{sm} - \frac{1}{N}\sum\limits_{i\in s_{msl}} a_i^*(sl)\mathbb{x}_{mi}\right)\mathbb{x}_{mi}}{\sum\limits_{i\in s_{msl}} \mathbb{x}_{mi}\mathbb{x}^t_{mi}q_{msli}a_i^*(sl)}\right] \tag{38}$$

with q_{msli} as known positive constant unrelated to $a_i^*(sl)$ and $\bar{\mathbb{X}}_{sm} = (\bar{y}^c_{1nsl},\ \bar{X}_1,\ \bar{X}_2,\ \ldots,\ \bar{X}_p)^t$ where

$$\bar{y}^c_{1nsl} = \left[\frac{1}{N} \sum_{i \in s_{nsl}} a_{1i}(sl) t_{1i} + \hat{\mathbb{B}}_{nsl} \left(\bar{\mathbb{X}} - \frac{1}{N} \sum_{i \in s_{nsl}} a_{1i}(sl) \mathbb{x}_i \right)^t \right] \tag{39}$$

with

$$\hat{\mathbb{B}}_{nsl} = \left(\sum_{i \in s_{nsl}} a_{1i}(sl) q_{nsli} \mathbb{x}_i \mathbb{x}_i^t \right)^{-1} \left(\sum_{i \in s_{nsl}} a_{1i}(sl) q_{nsli} \mathbb{x}_i t_{1i} \right)$$

with q_{nsli} as known positive constant unrelated to $a_{1i}(sl)$ and $\bar{\mathbb{X}}$.
Now, substituting the calibrated weights w_{mlli} in equation (37) and w_{msli} in equation (38) in to equation (32), the final proposed IST calibrated estimator T_{cm} based on sample size m at current wave becomes

$$T_{cm} = \left[\frac{1}{N} \sum_{i \in s_{mll}} a^*_i(ll) z_{2i} + \hat{\mathbb{B}}_{mll} \left(\bar{\mathbb{X}}_{lm} - \frac{1}{N} \sum_{i \in s_{mll}} a^*_i(ll) \mathbb{x}_{mi} \right)^t \right] - \left[\frac{1}{N} \sum_{i \in s_{msl}} a^*_i(sl) t_{2i} + \right.$$
$$\left. \hat{\mathbb{B}}_{msl} \left(\bar{\mathbb{X}}_{sm} - \frac{1}{N} \sum_{i \in s_{msl}} a^*_i(sl) \mathbb{x}_{mi} \right)^t \right]. \tag{40}$$

with,

$$\hat{\mathbb{B}}_{mll} = \left(\sum_{i \in s_{mll}} a^*_i(ll) q_{mlli} \mathbb{x}_{mi} \mathbb{x}_{mi}^t \right)^{-1} \left(\sum_{i \in s_{mll}} a^*_i(ll) q_{mlli} \mathbb{x}_{mi} z_{2i} \right)$$

and

$$\hat{\mathbb{B}}_{msl} = \left(\sum_{i \in s_{msl}} a^*_i(sl) q_{msli} \mathbb{x}_{mi} \mathbb{x}_{mi}^t \right)^{-1} \left(\sum_{i \in s_{msl}} a^*_i(sl) q_{msli} \mathbb{x}_{mi} t_{2i} \right).$$

15.3 Composite IST Successive Calibration Estimator

Considering the convex linear combination of the proposed IST Calibration estimators T_{cu} & T_{cm}, the final IST Calibration estimator for p-additional available auxiliary variables is proposed as

$$T_c = \phi_c T_{cu} + (1 - \phi_c) T_{cm} \tag{41}$$

where T_{cu} and T_{cm} are given in equation (30) and equation (40) respectively and $\phi_c \in [0,\ 1]$ is a scalar quantity to be chosen suitably.

16 Asymptotic Variance of IST Calibration estimator

This section is devoted to elaboration of asymptotic properties of the proposed IST calibration estimator T_c. As the estimator T_c depends on the estimators T_{cu} and T_{cm} given in equation (30) and equation (40) respectively, therefore first we discuss the asymptotic properties of T_{cu} and T_{cm}. From equation (31), it is observed that T_{cu} further depends on T_{cull} and T_{cusl}. Hence, following results due to Randles(1982), some theorem can be established for T_{cull} as

16.1 Theorem. *The asymptotic behaviour of the IST calibration estimator T_{cull} is same as that of*

$$T_{cull|B} = \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)z_{2i} + \left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)\mathbb{x}_i\right)^t \mathbb{B} \tag{42}$$

with

$$\mathbb{B} = \left(\sum_{i\in s_{ull}} q_{ulli}\mathbb{x}_i\mathbb{x}_i^t\right)^{-1}\left(\sum_{i\in s_{ull}} q_{ulli}\mathbb{x}_i z_{2i}\right) \tag{43}$$

Proof. Let us assume

$$T_{cull}(\alpha) = \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)z_{2i} + \left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)\mathbb{x}_i\right)^t \alpha \tag{44}$$

where α is assumed to be a p-dimentional vector.

Equation (44) shows $T_{cull}(\alpha)$ is the calibration estimator T_{cull} when $\hat{\mathbb{B}}_{ull}$ is replaced by a vector of variables α.

Therefore, the limiting mean of $T_{cull}(\alpha)$ when the actual parameter value is $\mathbb{B}$ (given in equation (43)) can be written as

$$\mu(\alpha) = \lim_{u_{ll}\to+\infty} E_B\left[T_{cull}(\alpha)\right] = \tilde{Z}_2 \tag{45}$$

where $\tilde{Z}_2$ is the limiting value of $\bar{Z}_2$ as $N \to \infty$.

Hence, by Randles(1982), the estimator T_{cull} has the same asymptotic behaviour as that of the estimator

$$T_{cull|B} = \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)z_{2i} + \left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{ull}} b_i^*(ll)\mathbb{x}_i\right)^t \mathbb{B} \tag{46}$$

16.2 Theorem. *The variance of the estimator $T_{cull|B}$ is given by*

$$V\left(T_{cull|B}\right) = \left[\frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\mathit{\Delta}^c_{ijll}\frac{z_{2i}}{(\pi_{ill})^c}\frac{z_{2j}}{(\pi_{jll})^c}\right] +$$

$$E_1\left[\frac{1}{N^2}\sum_{i\in(s_{nll})^c}\sum_{j\in(s_{nll})^c}\frac{\mathit{\Delta}_{ijll}|(s_{nll})^c}{(\pi'_{ill})^c(\pi'_{jll})^c}\frac{e_i}{\pi_{ill}|(s_{nll})^c}\frac{e_j}{\pi_{jll}|(s_{nll})^c}\right] \qquad (47)$$

where, $\mathbf{e}_i = z_{2i} - (\mathbb{x}_i)^t\mathbb{B}$ *and* E_1 *is the expectation under the design* d_1.

Proof. Since, the estimator $T_{cull|B}$ is unbiassed. Hence, its variance is given by

$$V(T_{cull|B}) = V_1\left(E_3\left[T_{cull|B}\right]\right) + E_1[V_3\left(T_{cull|B}\right)] \qquad (48)$$

where, E_1 and V_1 are the expectation and variance under the design d_1 respectively, and E_3 and V_3 represent the conditional expectation and conditional variance under design d_3 respectively.

$$\begin{aligned} V_1\left(E_3\left[T_{cull|B}\right]\right) =& V_1\left(\frac{1}{N}\sum_{i\in s_{ull}} b^*_i(ll)z_{2i}\right) \\ =& \frac{1}{N^2}\sum_{i\in s^c_n}\sum_{j\in s^c_n}\mathit{\Delta}_{ij}\frac{z_{2i}}{\pi_i}\frac{z_{2j}}{\pi_j} \end{aligned} \qquad (49)$$

Now,

$$\begin{aligned} E_1\left[V_3\left(T_{cull|B}\right)\right] =& E_1\left[V_3\left\{\left(\frac{1}{N}\sum_{i\in s_{ull}} b^*_i(ll)z_{2i}\right) + \left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\in s_{ull}} b^*_i(ll)\mathbb{x}_i\right)^t\mathbb{B}\right\}\right] \\ =& E_1\left[V_3\left(\frac{1}{N}\sum_{i\in s_{ull}} b^*_i(ll)z_{2i} - \frac{1}{N}\sum_{i\in s_{ull}} b^*_i(ll)\mathbb{x}_i\mathbb{B}\right)\right] \\ =& E_1\left[V_3\left(\frac{1}{N}\sum_{i\in s_{ull}} b^*_i(ll)e_i\right)\right] \\ =& E_1\left[\frac{1}{N^2}\sum_{i\in s_{ull}}\sum_{j\in s_{ull}}\frac{\mathit{\Delta}_{ij}}{\pi_i\pi_j}\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}\right] \end{aligned} \qquad (50)$$

Using equation (49) and equation (50) in equation (48), we get the expression for variance as in equation (47).

16.1 Remark. From Theorem 16.1 and Theorem 16.2, the estimator T_{cu} and T_{cm} are asymptotically unbiased and their asymptotic variances are given by

$$V\left(T_{cu}\right) = \frac{1}{N}\Bigg[\sum_{i\in U}\sum_{j\in U}\varDelta^{c}_{ijll}\frac{z_{2i}}{(\pi_{ill})^{c}}\frac{z_{2j}}{(\pi_{jll})^{c}} - \sum_{i\in U}\sum_{j\in U}\varDelta^{c}_{ijsl}\frac{t_{2i}}{(\pi_{isl})^{c}}\frac{t_{2j}}{(\pi_{jsl})^{c}}\Bigg] +$$

$$\frac{1}{N^2}E_1\Bigg[\sum_{i\in(s_{nll})^c}\sum_{j\in(s_{nll})^c}\frac{\varDelta_{ijll}|(s_{nll})^c}{(\pi'_{ill})^c(\pi'_{jll})^c}\frac{e_i}{\pi_{ill}|(s_{nll})^c}\frac{e_j}{\pi_{jll}|(s_{nll})^c} -$$

$$\sum_{i\in(s_{nsl})^c}\sum_{j\in(s_{nsl})^c}\frac{\varDelta_{ijsl}|(s_{nsl})^c}{(\pi'_{isl})^c(\pi'_{jsl})^c}\frac{\xi_i}{\pi_{isl}|(s_{nsl})^c}\frac{\xi_j}{\pi_{jsl}|(s_{nsl})^c}\Bigg] \tag{51}$$

Similarly,

$$V\left(T_{cm}\right) = \frac{1}{N}\Bigg[\sum_{i\in U}\sum_{j\in U}\varDelta'_{ijll}\frac{z_{2i}}{(\pi_{ill})^{c}}\frac{z_{2j}}{(\pi_{jll})^{c}} - \sum_{i\in U}\sum_{j\in U}\varDelta'_{ijsl}\frac{t_{2i}}{(\pi_{isl})^{c}}\frac{t_{2j}}{(\pi_{jsl})^{c}}\Bigg] +$$

$$\frac{1}{N^2}E_2\Bigg[\sum_{i\in s_{nll}}\sum_{j\in s_{nll}}\frac{\varDelta_{ijll}|s_{nll}}{\pi'_{ill}\pi'_{jll}}\frac{e_i}{\pi_{ill}|s_{nll}}\frac{e_j}{\pi_{jll}|s_{nll}} -$$

$$\sum_{i\in s_{nsl}}\sum_{j\in s_{nsl}}\frac{\varDelta_{ijsl}|s_{nsl}}{\pi'_{isl}\pi'_{jsl}}\frac{\xi_i}{\pi_{isl}|s_{nsl}}\frac{\xi_j}{\pi_{jsl}|s_{nsl}}\Bigg] \tag{52}$$

$\xi_i = t_{2i} - (\mathbb{x}_i)^t\mathbb{B}$ and E_2 is the expectation under the design d_2.

16.3 Theorem. *The asymptotic variance of proposed IST Calibration estimator for p-auxiliary variables is obtained as*

$$V\left(T_c\right) = \phi_c^2 V\left(T_{cu}\right) + \left(1-\phi_c\right)^2 V\left(T_{cm}\right) \tag{53}$$

where $V(T_{cu})$ *and* $V(T_{cm})$ *are given in equation (51) and equation (52) respectively*

Proof. The asymptotic variance of IST Calibration estimators T_c is given by

$$\begin{aligned} V\left(T_c\right) &= E\left(T_c\ -\ \bar{Y}_2\right)^2 \\ &= E\left[\phi_c\left(T_{cu} - \bar{Y}_2\right) + \left(1-\phi_c\right)\left(T_{cm} - \bar{Y}_2\right)\right]^2 \\ &= \phi_c^2 V\left(T_{cu}\right) + \left(1-\phi_c\right)^2 V\left(T_{cm}\right) + 2\phi_c\left(1-\phi_c\right) cov\left(T_{cu},\ T_{cm}\right) \end{aligned} \tag{54}$$

Since, T_{cu} and T_{cm} are based on two non-overlaping samples of sizes u and m respectively. So cov$(T_{cu},\ T_{cm}) = 0$. Substituting the values of $V\left(T_{cu}\right)$ and $V\left(T_{cm}\right)$ from the equation (51) and equation (52) respectively in the above equation (54), we have the expression for the asymptotic variance of the IST Calibration estimator T_c as in equation (53).

17 Optimum variance of IST Calibration estimator

It is to be noted that, $V\left(T_c\right)$ is a function of unknown constant ϕ_c. Hence, it is optimized with respect to ϕ_c and subsequently the optimum value of ϕ_c is obtained as

$$\phi_{c(opt.)} = \frac{V\left(T_{cm}\right)}{V\left(T_{cu}\right) + V\left(T_{cm}\right)} \tag{55}$$

Substituting $\phi_{c(opt.)}$ from equation (55) in equation (53), we get the optimum variance of the proposed IST Calibration estimator T_c as

$$V\left(T_c\right)_{opt.} = \frac{V\left(T_{cu}\right) \times V\left(T_{cm}\right)}{V\left(T_{cu}\right) + V\left(T_{cm}\right)} \tag{56}$$

18 Study under SRSWOR sampling design

In this section, we study the proposed IST calibration estimator under SRSWOR(Simple Random Sampling Without Replacement) sampling design. Therefore, we consider

$$\pi_i^{'} = \frac{n}{N}; \pi_{ij}^{'} = \frac{n(n-1)}{N(N-1)}$$

Because the sample s_n is drawn from U with SRSWOR of size n, it implies that the complement $s_n^c = U - s_n$ is a simple random sample without replacement of size $N-n$, therefore we have

$$\pi_i^{'c} = \frac{N-n}{N} \; ; \; \pi_{ij}^{'c} = \frac{(N-n)(N-n-1)}{N(N-1)}$$

Also, we suppose that the matched sample s_m is drawn from s_n with SRSWOR of size m so

$$\pi_{i|s_n} = \frac{m}{n} \; ; \; \pi_{ij|s_n} = \frac{m(m-1)}{n(n-1)}$$

Finally, the unmatched sample s_u is drawn from s_n^c with SRSWOR of size u. Thus we have

$$\pi_{i|s_n^c} = \frac{u}{N-u} \; ; \; \pi_{ij|s_n^c} = \frac{u(u-1)}{(N-n)(N-n-1)}$$

Now, based on sample of size u on current occasion, the proposed IST calibration estimator T_{cu} under SRSWOR sampling design is obtained as

$$T_{cu}^{s} = \left[\bar{z}_{2u} + \mathbb{B}_{ull}(\bar{\mathbb{X}} - \bar{\mathbb{x}}_u)^t\right] - \left[\bar{t}_{2u} + \mathbb{B}_{usl}(\bar{\mathbb{X}} - \bar{\mathbb{x}}_u)^t\right] \tag{57}$$

where, $\bar{\mathbb{x}}_u = (\bar{x}_{1u},\ \bar{x}_{2u},\ \ldots,\ \bar{x}_{pu})^t$.

Similarly, based on sample of size m on current occasion, the proposed IST calibration estimator T_{cm} under SRSWOR scheme is obtained as

$$T^s_{cm} = \left[\bar{z}_{2m} + \mathbb{B}_{mll}(\bar{\mathbb{X}}_{ll} - \bar{\mathbb{x}}_m)^t\right] - \left[\bar{t}_{2m} + \mathbb{B}_{msl}(\bar{\mathbb{X}}_{sl} - \bar{\mathbb{x}}_m)^t\right] \tag{58}$$

Where, $\bar{\mathbb{x}}_m = (\bar{y}_{1m},\ \bar{x}_{1m},\ \bar{x}_{2m},\ \ldots,\ \bar{x}_{pm})^t$, $\bar{\mathbb{X}}_{ll} = (\bar{y}^s_{1nll},\ \bar{X}_1,\ \bar{X}_2,\ \ldots,\ \bar{X}_p)^t$ and $\bar{\mathbb{X}}_{sl} = (\bar{y}^s_{1nsl},\ \bar{X}_1,\ \bar{X}_2,\ \ldots,\ \bar{X}_p)^t$ and

$$\bar{y}^s_{1nll} = \left[\bar{z}_{1n} + \mathbb{B}_{nll}(\bar{\mathbb{X}} - \bar{\mathbb{x}}_n)^t\right],\quad \bar{y}^s_{1nsl} = \left[\bar{t}_{1n} + \mathbb{B}_{nsl}(\bar{\mathbb{X}} - \bar{\mathbb{x}}_n)^t\right]$$

where, $\bar{\mathbb{x}}_n = (\bar{x}_{1n},\ \bar{x}_{2n},\ \ldots,\ \bar{x}_{pn})^t$.

Now, taking convex linear combination of the proposed IST Calibration estimators T^s_{cu} and T^s_{cm} for p-auxiliary variables under SRSWOR sampling design, we get

$$T^s_c = \phi^s_c T^s_{cu} + (1 - \phi^s_c)T^s_{cm} \tag{59}$$

where T^s_{cu} and T^s_{cm} are given in equation (57) and equation (58) respectively and $\phi^s_c \in [0,\ 1]$ is a scalar quantity to be chosen suitably.

18.1 Remark. Further if we assume, $q_{ulli} = q_{usli} = q_{mlli} = q_{msli} = 1$ in T_c, then the IST calibration estimator is becomes T^*_c given as

$$T^*_c = \phi^*_c T^*_{cu} + (1 - \phi^*_c)T^*_{cm}\ ;\ \phi^*_c \in [0,\ 1]$$

where,

$$T^*_{cu} = \left[\frac{1}{N}\sum_{i\epsilon s_{ull}} b^*_i(ll)z_{2i} + \hat{\mathbb{B}}^*_{ull}\left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\epsilon s_{ull}} b^*_i(ll)\mathbb{x}_i\right)^t\right] - \left[\frac{1}{N}\sum_{i\epsilon s_{usl}} b^*_i(sl)t_{2i} + \right.$$

$$\left.\hat{\mathbb{B}}^*_{usl}\left(\bar{\mathbb{X}} - \frac{1}{N}\sum_{i\epsilon s_{usl}} b^*_i(sl)\mathbb{x}_i\right)^t\right]$$

with

$$\hat{\mathbb{B}}^*_{ull} = \left(\sum_{i\epsilon s_{ull}} b^*_i(ll)\mathbb{x}_i\mathbb{x}^t_i\right)^{-1}\left(\sum_{i\epsilon s_{ull}} b^*_i(ll)\mathbb{x}_i z_{2i}\right),$$

$$\hat{\mathbb{B}}^*_{usl} = \left(\sum_{i\epsilon s_{usl}} b^*_i(sl)\mathbb{x}_i\mathbb{x}^t_i\right)^{-1}\left(\sum_{i\epsilon s_{usl}} b^*_i(sl)\mathbb{x}_i t_{2i}\right)$$

and

$$T^*_{cm} = \left[\frac{1}{N}\sum_{i\epsilon s_{mll}} a^*_i(ll)z_{2i} + \hat{\mathbb{B}}^*_{mll}\left(\bar{\mathbb{X}}_{lm} - \frac{1}{N}\sum_{i\epsilon s_{mll}} a^*_i(ll)\mathbb{x}_{mi}\right)^t\right] - \left[\frac{1}{N}\sum_{i\epsilon s_{msl}} a^*_i(sl)t_{2i} + \right.$$

$$\left.\hat{\mathbb{B}}^*_{msl}\left(\bar{\mathbb{X}}_{sm} - \frac{1}{N}\sum_{i\epsilon s_{msl}} a^*_i(sl)\mathbb{x}_{mi}\right)^t\right]$$

with,

$$\hat{\mathbb{B}}^*_{mll} = \left(\sum_{i\epsilon s_{mll}} a_i^*(ll)\mathbb{x}_{mi}\mathbb{x}_{mi}^t\right)^{-1}\left(\sum_{i\epsilon s_{mll}} a_i^*(ll)\mathbb{x}_{mi}z_{2i}\right),$$

$$\hat{\mathbb{B}}^*_{msl} = \left(\sum_{i\epsilon s_{msl}} a_i^*(sl)\mathbb{x}_{mi}\mathbb{x}_{mi}^t\right)^{-1}\left(\sum_{i\epsilon s_{msl}} a_i^*(sl)\mathbb{x}_{mi}t_{2i}\right).$$

19 Simulation Study

In order to reveal the behaviour of the proposed estimators and to compare them, a simulation study has been carried out. For simplicity, the simulation study has been carried out under SRSWOR (Simple Random Sampling Without Replacement).

For this purposes a natural population has been used from statistical abstracts of United States. The study and auxiliary variables for the considered population having $N = 51$ is described as:
y_1:Rate of abortion in 2007
y_2:Rate of abortion in 2008
t_1:Rate of residence in 2007
t_2:Rate of residence in 2008
x_1:Rate of abortion in 2000
x_2:Rate of abortion in 2005

We compare the IST calibration estimator with IST Horvitz-Thomson estimator under both the allocation designs. In order to do simulation with considered data $10,000$ independent samples have been generated under sampling in two wave under IST frame work. All samples are obtained under simple random sampling without replacement.

The performance of proposed IST calibration estimator T_c for $p = 1$ and $p = 2$ additional auxiliary variables has been evaluated in terms of Absolute relative bias (ARB) and Percent relative efficiency (PRE) under both Trappmann *et al.*(2014) as well as Perri *et al.*(2018) optimum allocation designs as discussed in Section (7.2).

$$ARB_k = \frac{1}{10000}\left|\frac{\sum\limits_{i=1}^{10000}\{\theta_i\}_k - \bar{Y}_2}{\bar{Y}_2}\right| \qquad \& \qquad PRE_{jk} = \frac{\sum\limits_{i=1}^{10000}[T_{Hi} - \bar{Y}_2]^2}{\sum\limits_{i=1}^{10000}[\{T_{ci}(p=j)\}_k - \bar{Y}_2]^2}$$

where $\theta \in \; \{T_H, \; T_c(p=j)\}$; for $j = 1, \, 2$ and $k = \begin{cases} 1 \text{ for Trappman } \textit{et al.}(2014) \text{ allocation design} \\ 2 \text{ for Perri } \textit{et al.}(2018) \text{ allocation design} \end{cases}$

The simulation results have been represented in various graphs (Figure 1 - Figure 8) for varying $\phi \; = \; 0.1, \; 0.2, \; \ldots, \; 0.9$ where $\phi \in \{\phi_H, \; \phi_c\}$.

20 Direct Method

In general the estimators under IST setup is less efficient than the estimators obtained using direct questioning. Hence, in order to identify the amount of loss, the proposed IST calibration estimator T_c has been compared with respect to corresponding estimator under direct questioning (without IST set up). The calibration estimator under direct questioning is proposed as

$$T_d = \phi_d T_{du} + (1-\phi_d) T_{dm} \; ; \; \phi_d \in [0, \; 1] \tag{60}$$

where,

$$T_{du} = \frac{1}{N} \sum_{i\epsilon s_u} w_{ui} y_{2i} = \frac{1}{N} \sum_{i\epsilon s_u} b_i^* y_{2i} + \hat{\mathbb{B}}_{ud} \left(\bar{\mathbb{X}} - \frac{1}{N} \sum_{i\epsilon s_u} b_i^* \mathbb{x}_i \right)^t \tag{61}$$

with

$$\hat{\mathbb{B}}_{ud} = \left(\sum_{i\epsilon s_u} b_i^* q_{ui} \mathbb{x}_i \mathbb{x}_i^t \right)^{-1} \left(\sum_{i\epsilon s_u} b_i^* q_{ui} \mathbb{x}_i y_{2i} \right)$$

and

$$T_{dm} = \frac{1}{N} \sum_{i\epsilon s_m} w_{mi} y_{2i} = \frac{1}{N} \sum_{i\epsilon s_m} a_i^* y_{2i} + \hat{\mathbb{B}}_{md} \left(\bar{\mathbb{X}}_m - \frac{1}{N} \sum_{i\epsilon s_m} a_i^* \mathbb{x}_{mi} \right)^t \tag{62}$$

with

$$\hat{\mathbb{B}}_{md} = \left(\sum_{i\epsilon s_m} a_i^* q_{mi} \mathbb{x}_{mi} \mathbb{x}_{mi}^t \right)^{-1} \left(\sum_{i\epsilon s_m} a_i^* q_{mi} \mathbb{x}_{mi} y_{2i} \right)$$

It has been observed that optimum allocation of LL and SL sample in IST under Perri *et al.*(2018) approach is more efficient than that of allocation by Trappmann *et al.*(2014) allocation. Hence, we have compared IST calibration estimator under Perri*et al.*(2018) optimum allocation with direct questioning method. The cases of $p = 1$ and $p = 2$ additional auxiliary variables have been discussed. The ratio of variance of IST calibration method with respect to direct questioning method have been computed via simulation as

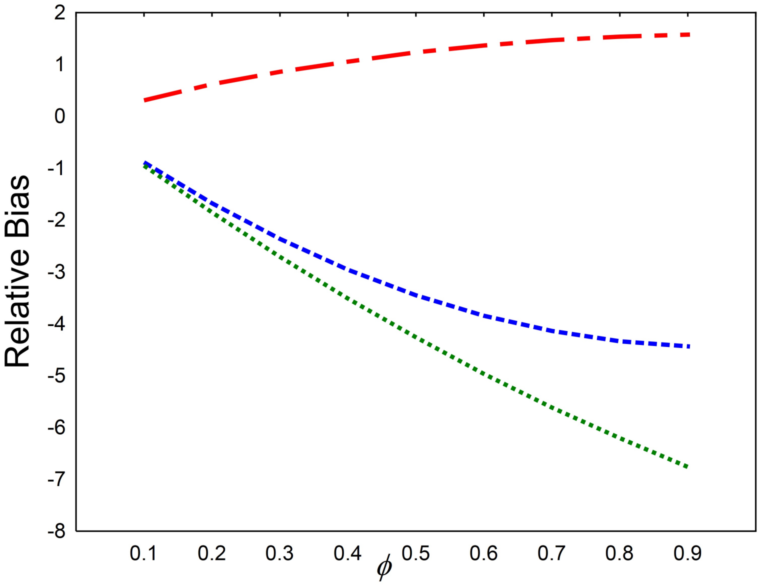


Figure 1: Relative Bias for k=1, n=30, u=14 & m=16

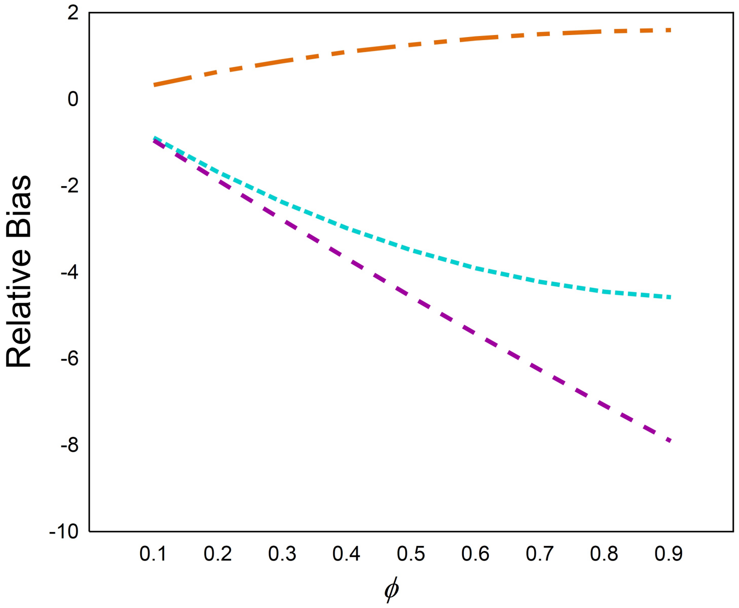


Figure 2: Relative Bias for k=1, n=30, u=12 & m=18

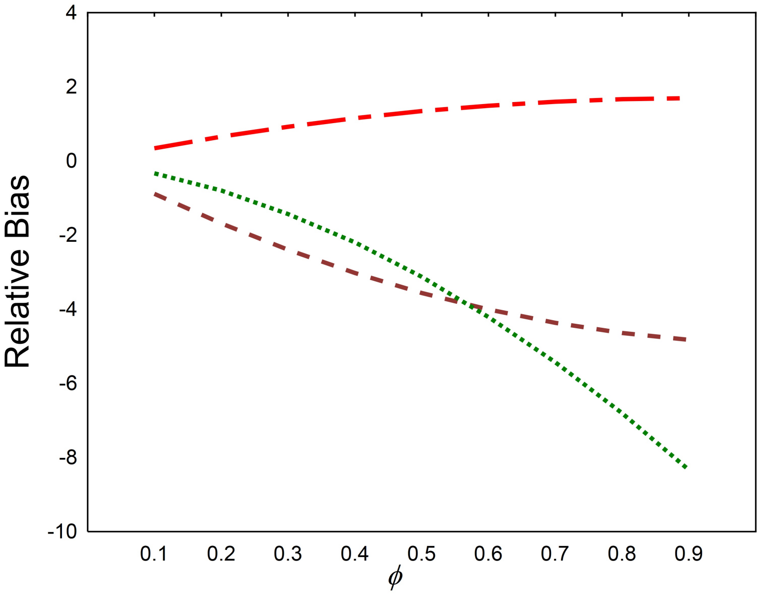


Figure 3: Relative Bias for k=2, n=30, u=8 & m=22

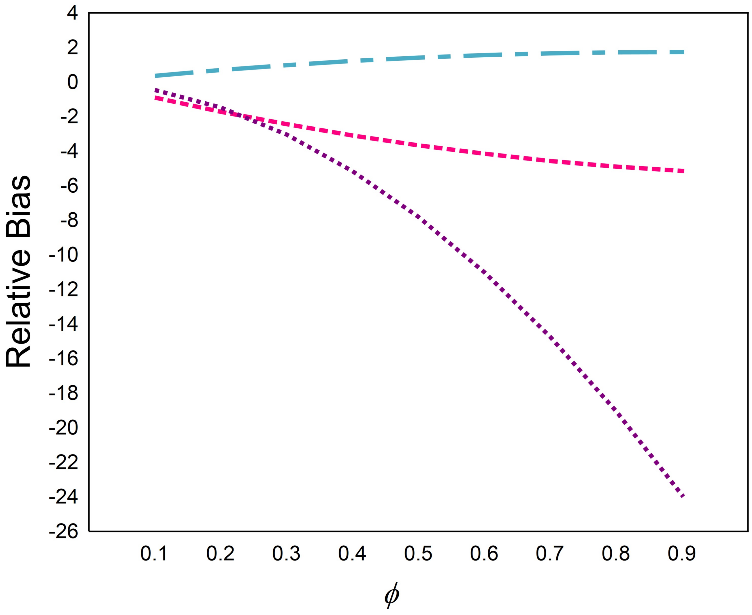


Figure 4: Relative Bias for k=2, n=30, u=4 & m=26

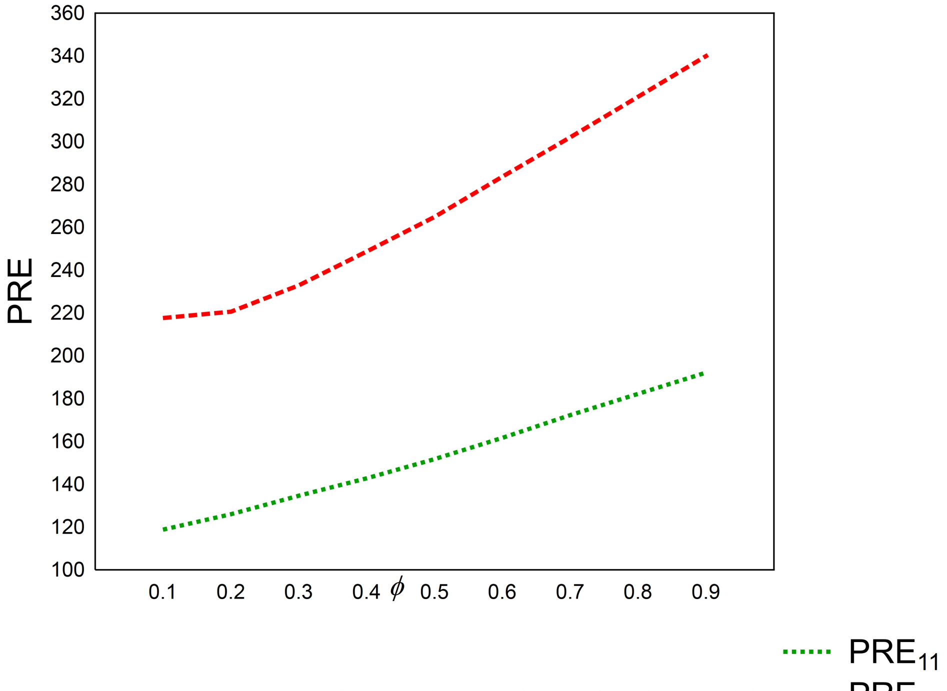


Figure 5: PRE for k=1, n=30, u=14 & m=16

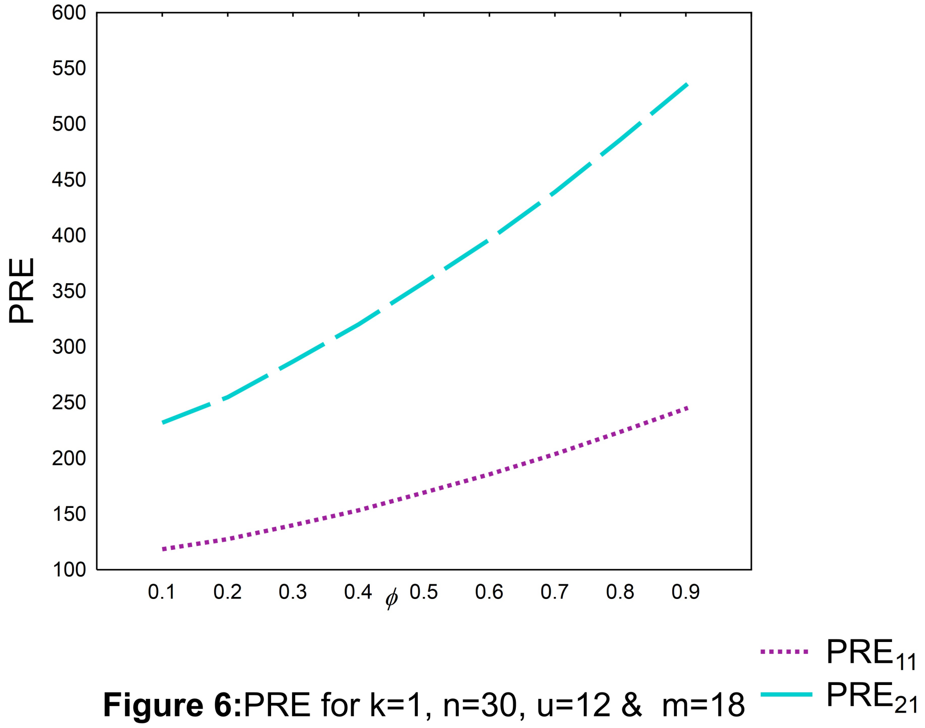


Figure 6:PRE for k=1, n=30, u=12 & m=18

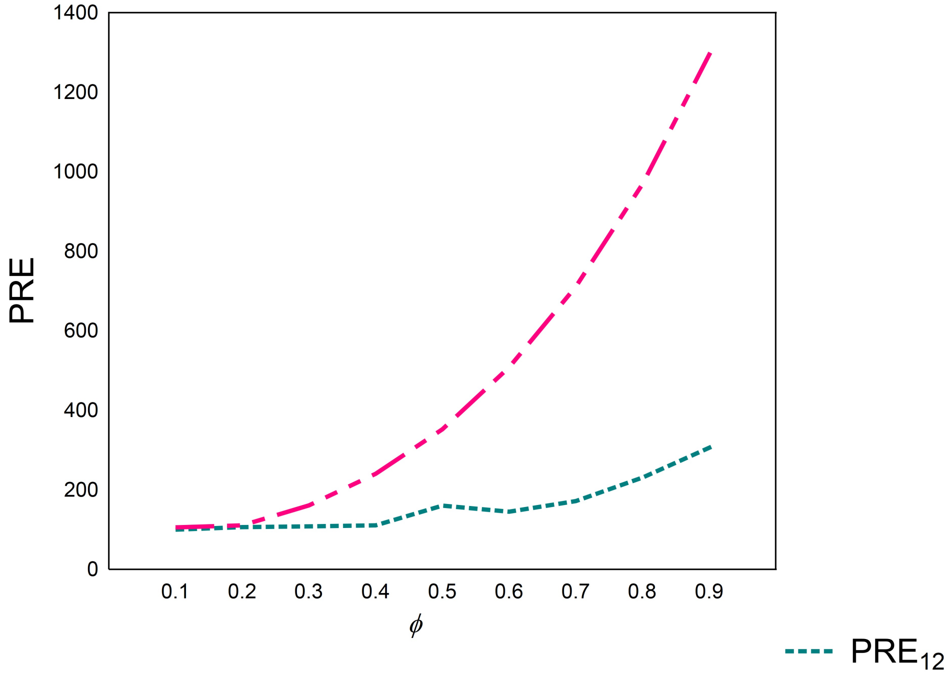


Figure 7: PRE for k=2, n=30, u=8 & m=22

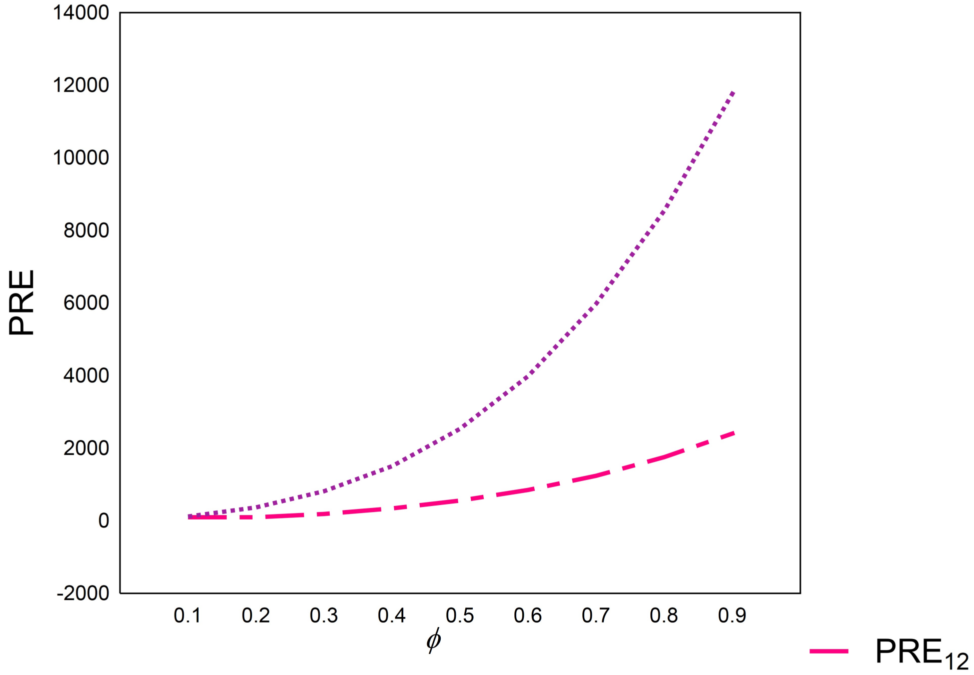


Figure 8: PRE for k=2, n=30, u=4 & m=26

$$Ratio_j = \frac{\sum\limits_{i=1}^{10000} \left[T_{di}(p=j) - \bar{Y}_2\right]^2}{\sum\limits_{i=1}^{10000} \left[T_{ci}(p=j) - \bar{Y}_2\right]^2} \; ; \; j = 1,\ 2$$

Ten thousand different replications of samples have been taken for simulation and the results have been demonstrated in different graphs(Figure 9 and Figure 10).

21 Discussion of Results

Following noteworthy results can be drawn from the simulation results shown in Figure 1 to Figure 8:

1. All the three considered estimators $T_H,\ T_c(p=1)$ & $T_c(p=2)$ have reasonable absolute relative bias under both Trappmann *et al.*(2014) as well as Perri *et al.*(2018) optimum allocation designs.
2. The IST calibrations estimators have less absolute relative bias than IST Horvitz-Thomson estimator. So, in terms of absolute relative bias IST calibration estimators are preferable over IST Horvitz Thomson estimator.
3. The percent relative efficiency confirm better behaviours of IST calibration estimators when compared to IST Horvitz-Thomson type estimator.
4. The percent relative efficiency increase with increase in number of auxiliary variable under both the allocation design considered. However, better PRE have been observed in case of optimum allocation design by Perri *et al.*(2018) than Trappmann *et al.*(2014) allocation design.
5. As enhanced PRE and less absolute relative bias has been observed in optimum allocation design. Hence, further study of comparison with direct questioning method have been worked out considering optimum allocation design only.
6. From simulation results in Figure 9 and Figure 10 it is clear that for some combination of parameters, the IST calibration estimator do not behave better than direct questioning method. This is the cost, we have to pay for using IST set up. However, if IST has not been used, then these might have created a situation that we may not have obtained honest responses as the issues under consideration are sensitve in nature. However, for few combination of constants we see that ratio is coming out to be more than one, this indicates that despite of IST, the estimators are so designed that it is coming out to be better than direct questioning method.

22 Conclusion

In order to deal with sensitive issues which are dynamic over time, the new methodology proposed have not only taken care of sensitivity of issues but

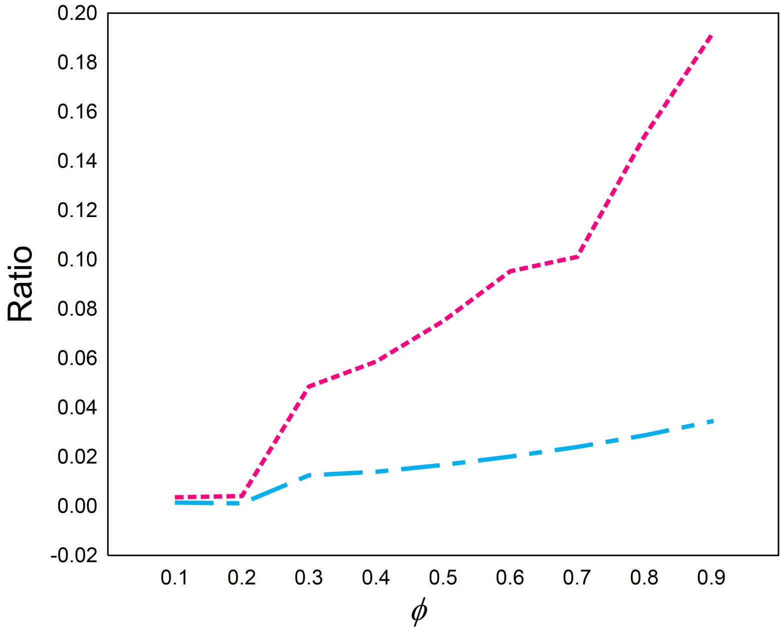


Figure 9: Ratio of variance for n=30, m=22 & u=8

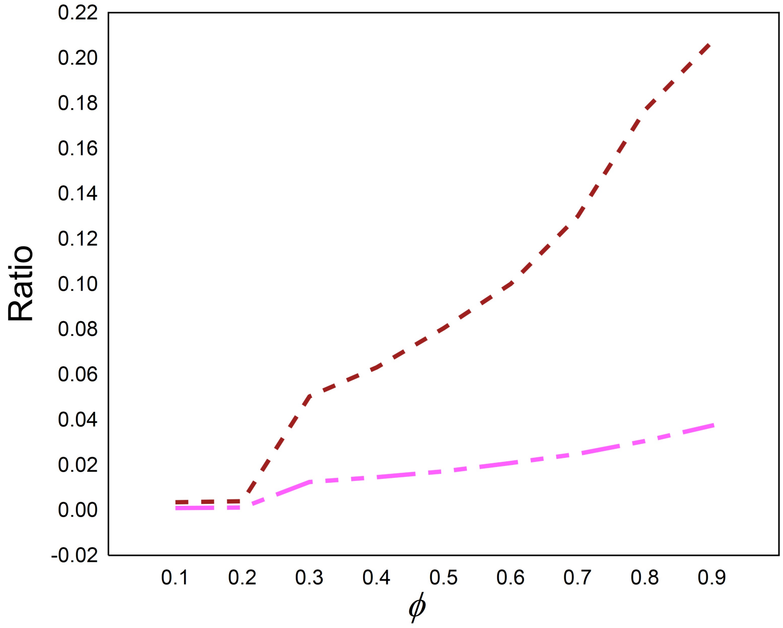


Figure 10: Ratio of variance for n=30, m=26 & u=4

also incorporates all the available auxiliary informations on both the waves in successive sampling. The concept of IST manages sensitivity of issues, however if bad samples are drawn, then that is managed by concept of calibration. Combining the two concepts together yields fruitful results to deal with sensitive issues over successive waves. The performance enhances as the number of auxiliary variables increases. The IST calibration estimators are coming out to be always efficienct than IST Horvitz-Thomson type estimators. Therefore, the methodology proposed can be recommended as efficient alternative with a wide number of desirable propertiers to be used by practitioners in this field.

23 Selected References

1. Arnab, R. and Singh, S. (2013). Estimation of mean of sensitive characteristics for successive sampling. *Comm. Statist.-Theo. Meth.*42:2499-2524.

2. Chaudhuri,A., and Christofides,T.C. (2013). Indirect questioning in sample surveys. *Springer-Verlag Berlin Heidelberg.*

3. Deville, J. C., Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*87:376-382.

4. Farrell, P.J., and Singh, S. (2002). Penalized chi-square distance function in survey sampling. *Joint statistical meetings-section on survey research methods,NY.*963-968.

5. Horvitz,D.G. and Thompson,D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*47:663-685.

6. Husssian,Z., Shabbir,N., and Shabbir,J. (2015). An Alternative Item Sum Technique for Improved Estimators of Population Mean in Sensitive Surveys. *Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics.* 46(91):DOI: 10.15672/HJMS.20159113160.

7. Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts, *Iowa Agri. Exp. Stat. Res. Bull.*304:1-104.

8. Kim, J.M., and Park,M. (2010). Calibration estimation in survey sampling. *Int. statist. Review.* 78:21-39.

9. Miller,J.D. (1984). A new survey technique for studying deviant behaviour. *Ph.D. thesis.* The George Washington University, Washington,DC.

10. Perri, P.F., Rueda García, M.d.M., Cobo Rodriguez, B. (2018). Multiple sensitive estimation and optimal sample size allocation in the item sum technique. *Biometrical Journal.*60:155-173.

11. Priyanka, K., Trisandhya, P., and Mittal, R. (2018). Dealing sensitive characters on successive occasions through a general class of estimators using scrambled response techniques. *Metron.*76(2):203-230.

12. Priyanka, K., and Trisandhya, P. (2019a). A Composite Class of Estimators using Scrambled Response Mechanism for Sensitive Population mean in Successive Sampling. *Comm. stat. Theory and Methods.* 48(4):1009-1032.

13. Priyanka, K. and Trisandhya, P.(2019b).The Item Sum Techniques for Quantitative Sensitive Estimation on Successive Occasions. Communications for Statistical Applications and Methods.26(2), 1-15, 2019.

14. Priyanka, K., Kumar,A. and Trisandhya, P.(2019c): Calibration Estimators for Quantitative Sensitive Mean Estimation Under Successive Sampling. Communications in Statistics- Theory and Methods. DOI:10.1080/03610926.2019.1649430.

15. Randles, R. (1982). On the asymptotic normality of statistics with estimated parameters. *Ann. Statist.*10:462-474.

16. Rueda,M., Martínez, S., Arcos,A., and Muñoz,J.F. (2009). Mean Estimation Under Successive Sampling with Calibration Estimators. *Comm. StatTheo and Meth.*38:808-827.

17. Rueda Garcia, M.d.M., Perri, P.F.,and Cobo Rodriguez, B.(2017). Advances in estimation by the item sum technique using Auxiliary information in complex surveys. *AStA Adv Stat Anal.*DOI:10.1007/s10182-017-0315-2.

18. Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.*33:99-119.

19. Trappmann, M., Krumpal, I., Kirchner,A., and Jann,B. (2014). Item sum:A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology.*2:58-77.

Random Forest for Classification and Regression

P. K. Meher
ICAR-IASRI, New Delhi
Prabina.Meher@icar.gov.in

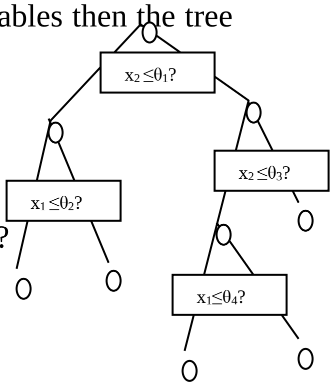
1. Introduction

Decision tree learning is a method commonly used to create a model that predicts the value of target variable based on several input variables. Decision trees are of two types; (i) classification tree (ii) regression tree. Classification tree analysis is there when the response variable is a class label and the regression tree analysis is there when the response variable takes the values of real number. A classification tree is obtained by asking an ordered sequence of questions, where the type of questions asked at each step in the sequence depends upon the answers required for the previous questions of the sequence. The sequence always terminates in a prediction of the class label attached to the observation. The starting point of a classification tree is called the root node and consists of the whole data set at the top of the tree. A node in a tree can be a terminal or non-terminal node. A non-terminal (or parent) node is a node that split into daughter nodes. A node that doesn’t split is called a terminal node and is assigned a class label. When an observation of unknown class is dropped down the tree and ends up at a terminal node, it is assigned to that class corresponding to the class label attached to that node. There may be more than one terminal node with the same class label. A single split tree with only two terminal nodes is called a stump. In case of binary splitted node, the split is determined by a Boolean condition on the value of a single variable, where the condition is either satisfied (“yes”) or not satisfied (“no”) by observed value of that variable. All the observations in the data set that have reached to a particular node and satisfy the condition for that variable drop down to one of the two daughter nodes and the remaining observations at that node that don’t satisfy the condition drop down to the other daughter node.

Let x_1 and x_2 be two variables and θ_i (i=1, 2, 3, 4) be any values of the variables then the tree is grown by asking following questions:

(1) Is $x_2 \leq \theta_1$? If the answer is yes, follow the left branch; if no follow the right branch.

(2) If the answer to question (1) is yes, then ask the next question: Is $x_1 \leq \theta_2$?



if the answer is yes, follow the left branch (terminal);

if no follow the right branch (terminal).

(3) If the answer to question (1) is no, ask the next question: Is $x_2 \leq \theta_3$?

(4) if the answer to (3) is yes, {then ask the next question: Is $x_1 \leq \theta_4$?

if the answer is yes, follow the left branch (terminal);

if no follow the right branch (terminal)}.

if the answer to (3) is no, it leads to the terminal node

2. Aspect of growing Tree

For growing a classification tree, following four aspects need to be discussed

- Choosing the Boolean conditions for splitting at each node
- Criterion to be used to split a parent node into its daughter nodes
- To decide a node to become a terminal node
- Assigning a class to a terminal node

3. Splitting strategies

In the splitting strategy the first two aspects of growing tree are discussed.

Number of possible splits

For continuous or ordinal variable, the total number of possible splits at a given node is one fewer than the number of its distinctly observed values. For nominal or categorical variable of *m* distinct categories, there will be 2^{m-1}-1dinstict splits at a particular node.

Node impurity function

To chose the best split among all variables, first chose the best split for a given variables by using measure of goodness of split. Let $\Pi_1,\ldots,\Pi_K$ be the K≥2 classes. For node τ, the node impurity function i(τ) is given as $i(\tau) = \phi(\, p(1|\tau),\ldots, p(K|\tau))$,Where $p(k/\tau)$ is an estimate of $P(Ð_K/ô)$ which is the conditional probability that an observation **X** is in $Ð_k$ given that it falls into node ô. The function ϕ will attain maxima at the point $(\frac{1}{K},\frac{1}{K},\ldots,\frac{1}{K})$ on the set of K-tuples of probabilities $(p_1,\ldots,p_K)$ and its sum is unity. In the two classes case (K=2), these condition reduces to a symmetric $\phi(p)$ maximized at the point p=1/2. One such function is the entropy function,

$$i(\tau) = -\sum_{k=1}^{K} p(k|\tau)\ \log p(k|\tau)$$

and for binary classes it reduces to $i(\tau) = -p\ \log p - (1-p)\ \log(1-p)$

Choosing best split for a variable

Let at node ô, after applying split *s*, a portion p_l goes to the daughter node $ô_l$ and the remaining portion p_r goes to the right daughter node $ô_r$. Then the goodness of split *s* at nod ô is the reduction in impurity gained by splitting the parent node ô in to its daughter nodes $ô_l$ and $ô_r$, which is given by $\Delta i(\tau) = i(\tau) - p_l\, i(\tau_l) - p_r\, i(\tau_r)$

For example, consider a data set having the response variable y that has two values 0 and 1 and suppose one of the possible split of the input variables x_j is $x_j \leq c$ vs. $x_j > c$, where c is some values of x_j. Then a 2×2 table can be prepared as follows:

	1	0	Row total
$x_j \leq c$	n_{11}	n_{12}	$n_{1.}$
$x_j > c$	n_{21}	n_{22}	$n_{2.}$
Column Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Now for the parent node ô, p_l=($n_{.1}$ / $n_{..}$) and p_r=($n_{.2}$ / $n_{..}$) so the impurity function at the parent node will be

$$i(\tau) = -\left(\frac{n_{.1}}{n_{..}}\right)\log_e\left(\frac{n_{.1}}{n_{..}}\right) - \left(\frac{n_{.2}}{n_{..}}\right)\log_e\left(\frac{n_{.2}}{n_{..}}\right)$$

Now for the daughter nodes $ô_l$ and $ô_r$, for $x_j \leq c$, p_l=(n_{11} / $n_{1.}$) and p_r=(n_{12} / $n_{1.}$) and for $x_j >$ c, p_l=(n_{21} / $n_{2.}$) and p_r=(n_{22} / $n_{2.}$). Then the impurity function at the daughter nodes will be

$$i(\tau_l) = -\left(\frac{n_{11}}{n_{1.}}\right).\log_e\left(\frac{n_{11}}{n_{1.}}\right) - \left(\frac{n_{12}}{n_{1.}}\right).\log_e\left(\frac{n_{12}}{n_{1.}}\right)$$

$$i(\tau_r) = -\left(\frac{n_{21}}{n_{2.}}\right).\log_e\left(\frac{n_{21}}{n_{2.}}\right) - \left(\frac{n_{22}}{n_{2.}}\right).\log_e\left(\frac{n_{22}}{n_{2.}}\right)$$

and the best split for the single variable x_j is the one that has largest value of $\Delta i(s, \tau)$ over all s ª S_j, the set of all possible split for x_j .

Choosing best split at a node

A tree starts with the root node, which consists of all observation. By using the goodness-of-fit criterion for a single variable, the best split at the root node for each of the variables x_1 to

x_r can be found. The best split *s* at the root node is then the one that has the largest value of $\Delta i(s, \tau)$ over all *r* single-variable best splits at that node.

4. Choosing terminal node

A node can be declared as a terminal node if it fails to be larger than certain predetermined size; that is , if n(ô) ≤ n_{min} , where n(ô) is the number of observations in node ô and n_{min} is some previously assumed minimum size of a node. The terminal node act as a break on the tree growth, the larger the value of n_{min}, the more severe the break. In another way a node can be declared as a terminal node if the largest goodness-of-fit value at that node is smaller than a certain predetermined limit. However, these stooping rules are not fruitful in reality. A better approach is to let the tree grow to saturation and then prune it back (Breiman *et al.* 1984).

5. Associating a class with the terminal node

Suppose at a terminal node ô there are n(ô) observation of which n_k(ô) are from class Π_k, k=1,…,K. then the class which corresponds to the largest of the {n_k(ô)} is assigned to ô. This is called plurality rule and it can be easily obtained from the Bayes's rule classifier, where the node ô can be assigned to the class Π_i if

$$p(\Pi_i|\tau) = \max_{1\le k\le K} p(\Pi_k|\tau)$$

Let p(ô^a Π_i) = p_i , (i=1,…,K), be the prior probability of the nod ô belonging to different classes i.e., p_i = n_i(ô)|n(ô)) and let p_i(ô) = p(ô|Π_i) be the probability distribution function of observations in node ô belonging to class Π_i. then the posterior probability of that node ô will be assigned the class Π_i is given by

$$p(\Pi_i|\tau) = \frac{p_i(\tau).p_i}{\sum_{k=1}^{K} p_k(\tau).p_k}$$

The Bayes's rule classifier for K classes assigns ô to that class with the highest posterior probability. Since the denominator is fixed for all the classes, the node ô will be assigned to the class Π_i if

$$p(\Pi_i|\tau) = \max_{1\le k\le K} p(\Pi_k|\tau)$$

6. Ensemble of classifiers

A well-known method of building classification systems is to build multiple classifiers, each from a subset of the original training set, such that the final classification decision is aggregated from all classifiers' decisions. This method is called the *classifier ensemble* method (Buhlmann et.al. 2004). For example, five classifiers could be built independently using five different subsets of the original training set. These five classifiers would produce five predictions of the class label for each new record, and the class with a plurality of votes would be the prediction of the entire ensemble. It is also possible to extend this simple voting scheme so that each individual classifier prediction is given a weight, perhaps based on its test accuracy. The overall prediction becomes the plurality of the weighted votes.

Classifiers in an ensemble can all have the same type, or they can be of different types. For example, an ensemble with three classifiers can consist of three decision trees, or it can consist of a decision tree, a neural network (Kantardzic, 2003), and a Bayesian network (Dunham, 2003). Both kinds of ensembles are known to perform better than single classifiers. The variance between classifiers is reduced in the case of classifiers of the same type, and the bias between classifiers is reduced for ensembles with different types of classifiers. The classification models of ensembles for both kinds are, therefore, more representative of the data than a single classifier. In other words, having multiple strong classifiers each built from a different sample of the dataset leads to a final classification decision with higher accuracy than a single classifier.

Generating the datasets used for training the classifiers in an ensemble can be done by different methods such as bootstrap sampling (bagging) (Breiman, 1994), and boosting (Freund and Schapire, 1996). Suppose that a dataset contains n records, each with m attributes. Bootstrap sampling or bagging generates the datasets each of size n by randomly sampling the records with replacement. Hence the training dataset for each tree contains multiple copies of some of the original records. Boosting maintains weights for records in the training set, such that these weights are updated after each classifier is trained according to the difficulty of classifying the current set of records. The weights are then used to derive the new sampling for the dataset.

6.1. Random Forest

Bagging (Bootstrap aggregating) was the first procedure that successfully combined the ensemble of tree classifiers to improve the performance over a single classifier (Breiman, 1996b). In bagging randomization is introduced only while selecting the data set on which each tree is grown. Random forest (Breiman, 2001) is an extension of this bagging procedure where another source of randomization is introduced by choosing a subset of m variables at each node and node is split on the basis of best split.

Let $L = \{(\mathbf{x}_i, y_i), i = 1, 2, ..., n\}$ is the learning data set where y_i is the response variable and it takes values from K classes and there are p variables in the data set. Random forest consists of ensemble of B classifiers $h_1(\mathbf{x}), h_1(\mathbf{x}), ..., h_B(\mathbf{x})$, where each classifier is constructed upon a bootstrap replica of the learning data set, by selecting randomly selecting a subset of m variables out of p variables and the best split is determined on the basis of m selected variables using gini index.. Each classifier votes for one of the classes for each test instances and test instance is classified by the label of winning class. As the individual trees are constructed upon a bootstrap replication, there is on an average 36.8% of instances are not playing any role in the construction of the tree. These instances are called out of bag (OOB) instances. These OOB instances are the source of data used in the random forest for estimating the classification error and to evaluate the performance of the random forest. Random forests are computationally very efficient and offer good prediction accuracy and are less sensitive to noisy data.

Some features of RF

Let (**x**, y) denote the learning instances having n number of observations where each vector of attributes **x** is labeled with class y_j, (j=1,2,…,c). The correct class is denoted by y. $p(y_j)$ is the probability of class y_j . denote the set of OOB instances for classifier h_b as O_b. Let $Q(\mathbf{x}, y_j)$ be the OOB proportion of votes for class y_j for input vector **x.**

$$Q(\mathbf{x}, y_j) = \frac{\sum_{b=1}^{B} I(h_b(\mathbf{x}) = y_j; (\mathbf{x}, y) \in O_b)}{\sum_{b=1}^{B} I(h_b(\mathbf{x}); (\mathbf{x}, y) \in O_b)}$$

The Margin function, strength and Correlation between classifiers in a RF is defined as follow.

Margin function- The “margin function” measures the extent to which the average vote for right class y exceeds the average vote for any other class. The margin function of the labeled observation (**x,** y) is $m(\mathbf{x}, \mathrm{y}) = P(h(\mathbf{x}) = y) - \max_{\substack{j=1 \\ j \neq y}}^{c} P(h(\mathbf{x}) = y_j)$. If m(x, y)>0, then h(**x**) correctly classifies y. h(**x**) denote a classifier that predict the label y for an observation **x.**

Strength- It is defined as the expected margin, and is computed as the average over the training set.

$$s = \frac{1}{n} \sum_{i=1}^{n} \left(Q(\mathbf{x}_i, y) - \max_{\substack{j=1 \\ j \neq y}}^{c} Q(\mathbf{x}_i, y_j) \right), where$$

$$Q(\mathbf{x}_i, y_j) = \frac{\sum_{b=1}^{B} I(h_b(\mathbf{x}) = y_j ; (\mathbf{x}, y) \in O_b)}{\sum_{b=1}^{B} I(h_b(\mathbf{x}) ; (\mathbf{x}, y) \in O_b)}$$

$$Q(\mathbf{x}_i, y) = \frac{\sum_{b=1}^{B} I(h_b(\mathbf{x}) = y ; (\mathbf{x}, y) \in O_b)}{\sum_{b=1}^{B} I(h_b(\mathbf{x}) ; (\mathbf{x}, y) \in O_b)}$$

where I(.) is the indicator function

7. Advantages

- People could understand and interpret easily after brief explanation
- Many data analysis techniques require data normalization, creation of dummy variable etc. but it requires little data preparation.
- Generally the techniques are specialized in analyzing data set having only one type of variable, but it handles both numerical and categorical data.
- Performs well with large data in a short time

Suggested Reading:

1. Kass, G. V. (1980). An explanatory technique for investigating large quantities of categorical data, *Applied Statistics,* **29**, 119-127.
2. Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and regression trees*, Boca Raton, FL: Wadsworth.
3. Zhang, H. and Singer, B. (1999). *Recursive portioning in the health sciences*, New York: Springer.

4. Izenman, A. J. (2008). *Modern multivariate statistical techniques; regression, classification and manifold learning*, New York: Springer.

Introduction to Particle Swarm Optimization

Santosha Rathod[1] and Mrinmoy Ray[2]
1-ICAR- IIRR, Hyderabad
2- ICAR- IASRI, New Delhi
santosha.rathod@icar.gov.in

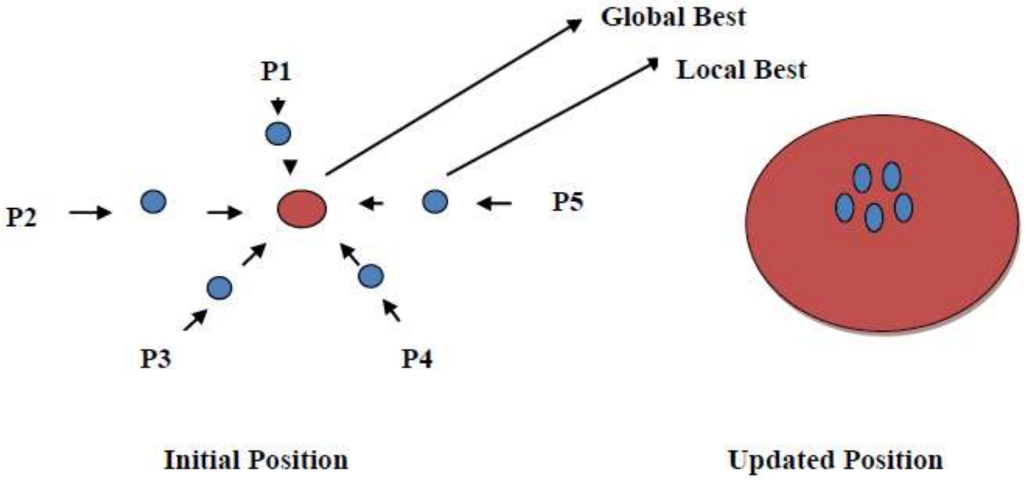
Introduction:

Particle swarm optimization (PSO) is a nature inspired evolutionary optimization technique to solve computationally hard or difficult optimization problems. It is a robust stochastic optimization technique based on the movement and intelligence of swarms. It was developed by James Knnedy and Russ Eberhart in 1995 based on the social behaviour of biological organisms that move in groups (swarms) such as birds and fishes. It has been applied successfully in wide variety of search and optimization problems by abstracting the working mechanism of natural phenomenon. Since PSO is a population-based (swarm) evolutionary algorithms, which has some similarities with GA. However, a fundamental difference between these paradigms is that the evolutionary algorithms are based on natural evolution concepts i.e. based on a competitive philosophy, it means only the fittest individuals tends to survive. Conversely, PSO incorporates a cooperative approach to solve a problem, given that all individuals (particles), which can survive, change themselves over time and one particle's successful adaptation is shared and reflected in the performance of its neighbours.

The basic element of PSO is a particle, which can fly throughout search space towards an optimum by using its own information as well as the information provided by other particles comprising its neighbourhood. In PSO, a swarm of n particles (individuals) communicate either directly or indirectly with one another using search directions (gradients). The algorithm adopted was a 'set of particles' flying over a search space to locate global optimum. During an iteration of PSO, each particle updates according to its previous experience and experience of its neighbours.

PSO Vectors:

X vector: Current location (current position) of the particle in search space, **P vector (pbest):** Location of best solution found so far by the particle and **V vector:** Gradient (direction) for which particle will travel in, if undisturbed. All these vectors are continuously updated.



Let, $A \subset R^n$be search space and the swarm is defined as a set $S = \{X_1, X_2, \dots, X_M\}$ of M particles (candidate solution), where M is a user-defined parameter of the algorithm. Then i^{th} particle dimension of d is defined as $X_i = (X_{i1}, \dots, X_{id})^T$, $i = 1,2, \dots, M$. Each particle is a potential solution to a problem, characterized by three quantities: velocity $V_i = (V_{i1}, \dots, V_{id})^T$, current position $X_i = (X_{i1}, \dots, X_{id})^T$and personal best position $pbest_i = (pbest_{i1}, \dots, pbest_{id})^T$. Let, t denote current iteration and $gbest$ denote its global best position achieved so far by any of its particles. Initially, swarm is randomly dispersed within search space and random velocity is assigned to each particle. Particles interact with one another by sharing information to discover optimal solution. Each particle moves in the direction of its personal best position ($pbest$) and its global best position ($gbest$). To search optimal solution, each particle changes its velocity according to the cognitive and social parts given by:

$$V_{ij}(t+1) = w(t)V_{ij}(t) + c_1R_1\left[pbest_{ij}(t) - X_{ij}(t)\right] + c_2R_2\left[gbest_j(t) - X_{ij}(t)\right]$$

Where, $i = 1,2, \dots, M$ and $j = 1,2, \dots, d$. However, in case of swarm explosion effect, corresponding velocity component is restricted to following closest velocity bound:

$V_{ij}(t+1) = -V_{max}$ if $V_{ij}(t+1) < -V_{max}$

$= V_{max}$ If, $V_{ij}(t+1) > V_{max}$

After updating its velocity, each particle moves to a new potential solution by updating its position as follows

$X_{ij}(t+1) = \ X_{min}$ if $X_{ij}(t+1) < X_{min}$

$= X_{ij}(t)+\beta V_{ij}(t+1)$, if $X_{min} \leq X_{ij}(t+1) \leq X_{max}$

$= X_{max}$, if $X_{ij}(t+1) > X_{max}$

Where, $i = 1,2,\dots,M$; $j = 1,2,\dots,d$. In the above equations V_{ij} , X_{ij} and $pbest_{ij}$ are respectively velocity, current position and personal best position of particle i on the j^{th}dimension, and $gbest_j$ is the j^{th}dimension global best position achieved so far among all particles at iteration t. R_1 and $R_2$are random values, which are mutually independent and uniformly distributed over $[0,1]$, β is a constraint factor used to control velocity weight, whose value is usually set equal to 1. Positive constants c_1 and c_2 are usually called “acceleration factors”. Factor c_1 is sometimes referred to as “cognitive” parameter, while c_1 is referred to as “social” parameter. Inertia weight at iteration t is $w(t)$ and is used to balance global exploration and local exploitation. This can be determined by:

$$w(t) = w_{up} - \left(w_{up} - w_{low}\right)t/T_{max}$$

Where,t is current iteration number, w_{up} and w_{low} are desirable lower and upper limits of w and T_{max} is maximum number of iterations.

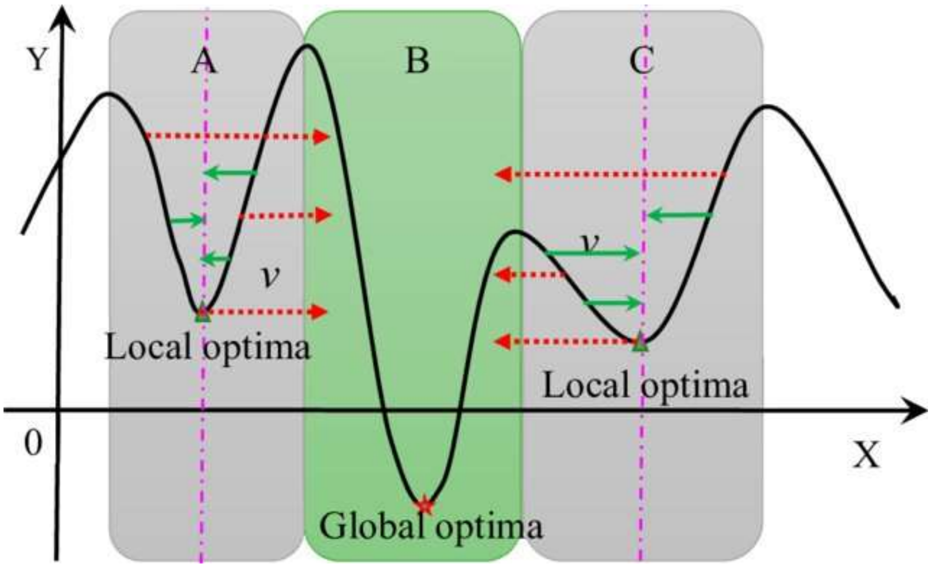
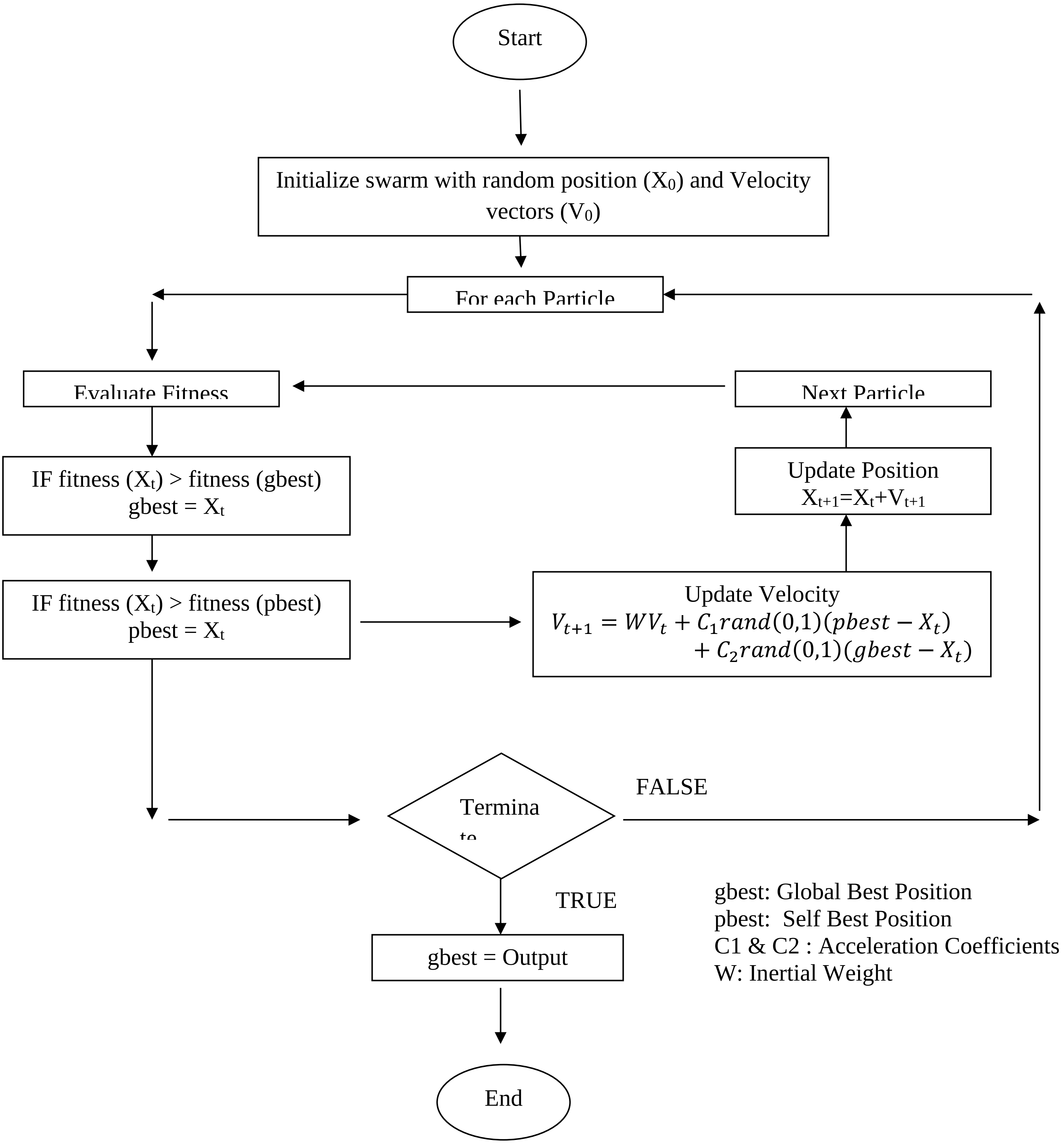


Fig.: Schematic diagram of particles' velocity.

Frame work of PSO:


Start
Initialize swarm with random position (X_0) and Velocity vectors (V_0)
For each Particle
Evaluate Fitness
Next Particle
IF fitness (X_t) > fitness (gbest)
gbest = X_t
Update Position
$X_{t+1}=X_t+V_{t+1}$
IF fitness (X_t) > fitness (pbest)
pbest = X_t
Update Velocity
$V_{t+1} = WV_t + C_1 rand(0,1)(pbest - X_t) + C_2 rand(0,1)(gbest - X_t)$
Termina
te
FALSE
TRUE
gbest = Output
End
gbest: Global Best Position
pbest: Self Best Position
C1 & C2 : Acceleration Coefficients
W: Inertial Weight

Algorithm Implementation:

Step 1: Initialize the parameters: initialize the position and speed of the particle to random numbers in the search space.

Step 2. Evaluate the particle's position: use a fitness function to evaluate each particle's position.

Step 3. Make a comparison between: (1) compare the fitness value of step 2 with the particle's personal best value pbest, and make the best value become the newest pbest; (2) compare the particle's fitness value with the global best value gbest, and the best one becomes gbest.

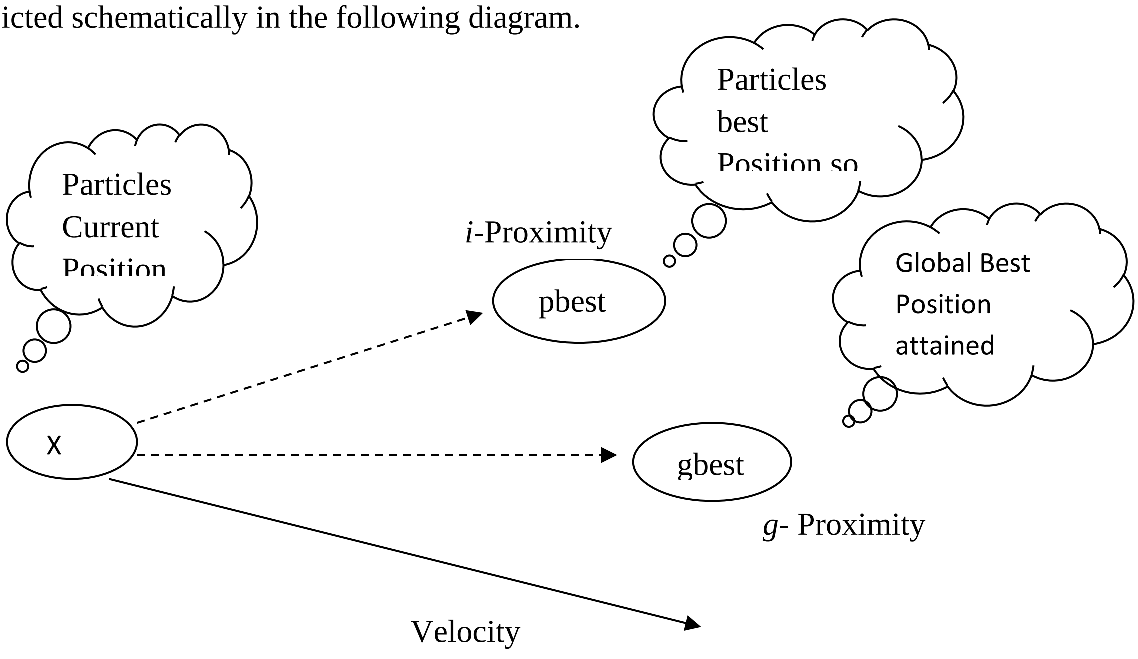
Step 4. Update the particle: Update the particle's speed and position.

Step 5. The termination conditions of iteration: circulate to step 2 until it satisfies the termination conditions, generally when the fitness value is optimal, or reaches the maximum iterations.

The basic concept of PSO lies in accelerating each particle towards the best position found by it so far (pbest) and the global best position (gbest) obtained so far by any particle, with a random weighted acceleration at each time step. This is done by simply adding V vector to X vector to get another X vector. $X_{i+1} = X_i + V_i$. Once, the particle computes the new X_i, it then evaluates its new location. IF X-Fitness is better than P-Fitness, then pbest=X_i and P-Fitness = X-Fitness

Psychosocial compromise:

Each particle updates its new position by compromising its local best towards the global best as depicted schematically in the following diagram.


Particles best Position so
Particles Current Position
i-Proximity
pbest
Global Best Position attained
X
gbest
g- Proximity
Velocity

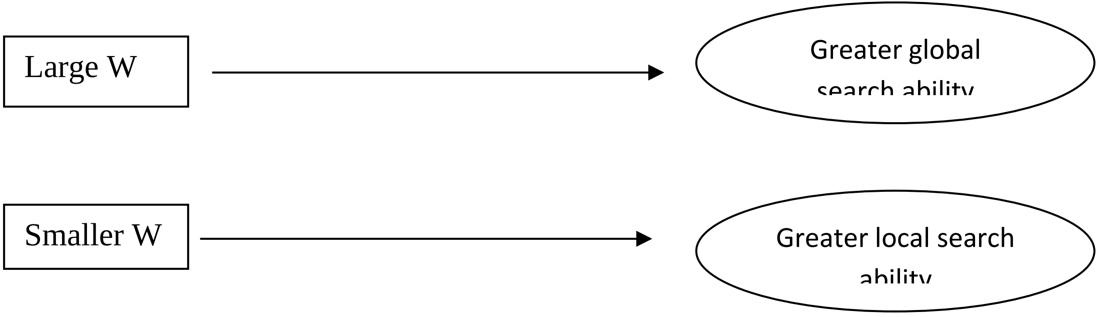
$$Position\ change\ is\ V_{t+1} = WV_t + C_1 rand(0,1)(pbest - X_t) + C_2 rand(0,1)(gbest - X_t)$$

User defined parameters:

Initial parameters such as swarm size, position of particles, velocity of particles and maximum number of iterations; and control parameters such as swarm size, inertial weight, acceleration coefficients C_1 and C_2 and number of iterations are very much important to begin with optimization algorithm. One has to define them in such a way that obtained parameter error should be less then target error.

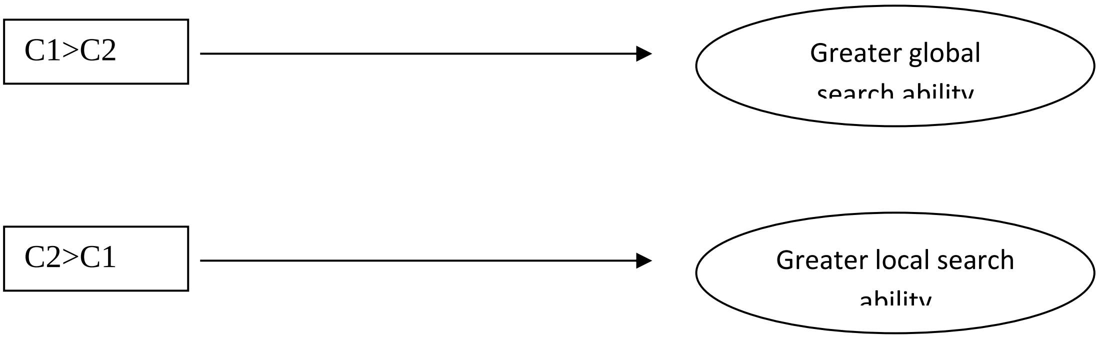
Innertial weight (W):

A large inertia weight (W) facilitates a global search while a small inertia weight facilitates a local search.



Acceleration coefficients:

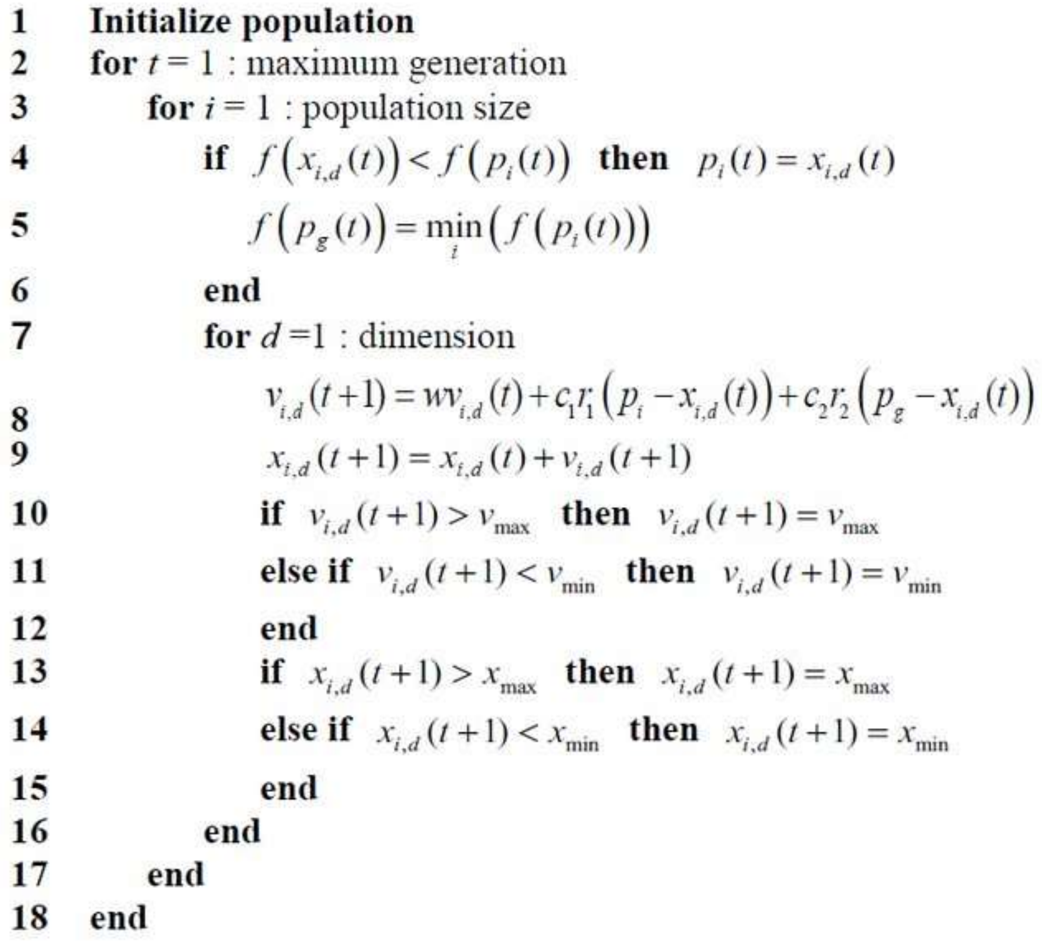
An acceleration coefficient determines the inclination of search, greater the C1, greater will be the global search ability, greater the C2, greater will be the local search ability.



Pseudo code of PSO:

```
For each particle
{
    Initialize particle
}
Do until maximum iterations or minimum error criteria
{
    For each particle
    {
        Calculate Data fitness value
        If the fitness value is better than pBest
        {
            Set pBest = current fitness value
        }
        If pBest is better than gBest
        {
            Set gBest = pBest
        }
    }
        For each particle
    {
        Calculate particle Velocity
        Use gBest and Velocity to update particle Data
    }
 }
```

Pseudocode in mathematical representation:



```
Initialize population
for $t = 1$ : maximum generation
    for $i = 1$ : population size
        if $f\left(x_{i,d}(t)\right) < f\left(p_i(t)\right)$ then $p_i(t) = x_{i,d}(t)$
            $f\left(p_g(t)\right) = \min_i\left(f\left(p_i(t)\right)\right)$
        end
        for $d = 1$ : dimension
            $v_{i,d}(t+1) = wv_{i,d}(t) + c_1r_1\left(p_i - x_{i,d}(t)\right) + c_2r_2\left(p_g - x_{i,d}(t)\right)$
            $x_{i,d}(t+1) = x_{i,d}(t) + v_{i,d}(t+1)$
            if $v_{i,d}(t+1) > v_{\max}$ then $v_{i,d}(t+1) = v_{\max}$
            else if $v_{i,d}(t+1) < v_{\min}$ then $v_{i,d}(t+1) = v_{\min}$
            end
            if $x_{i,d}(t+1) > x_{\max}$ then $x_{i,d}(t+1) = x_{\max}$
            else if $x_{i,d}(t+1) < x_{\min}$ then $x_{i,d}(t+1) = x_{\min}$
            end
        end
    end
end
```

Numerical Example 1:

[Reference: Mohanty, P. (2018). NTPL online certification course on selected topics on decision modelling, Particle Swarm Optimization, IIT Khargapur. https://www.youtube.com/watch?v=uwXFnzWaCY0]

Consider a maximization problem for maximization of the function $f(x) = 1 + 2x - x^2$

Let us consider the control parameters W=0.70, C1=0.20, C2=0.60 and n=5 (Swarm particle). Consider, random numbers used for updating velocity of particle be

r1 = [0.4657, 0.8956, 0.3877, 0.4902, 0.5039]

r2 = [0.5319, 0.8185, 0.8331, 0.7677, 0.1708]

Note: We keep the random numbers fixed for all the iterations throughout and each random number is corresponding to each particle.

Initialization of swarm particles: We initialize fitness of all the particles as zeros;

Current position of all the particles as;

Cp(0)=10*[r1-0.5]

Cp(0)=10*{[0.4657, 0.8956, 0.3877, 0.4902, 0.5039]-0.5}

So, Cp(0)=[-0.3425, 3.9558, -1.128, -0.0981, 0.0385]

Note: Multiplied by 10 to initialize at least some particles to be >1 and subtracted 0.5 sides to generate both positive and negative random numbers.

Initialization of velocity:

V(0)=r2-0.5

V(0)={[0.5319, 0.8185, 0.8331, 0.7677, 0.1708]-0.5}

We get,

V(0)=[0.0319, 0.3185, 0.3331, 0.2677, -0.3292]

Note: one should see that velocity should not be too high or too low.

Current position and current fitness:

Iteration 1:

Current position (Cp) of each particle is what we initialize

Cp(1)= Cp(0)= [-0.3425, 3.9558, -1.128, -0.0981, 0.0385]

Current velocity V(1)=V(0)

=[0.0319, 0.3185, 0.3331, 0.2677, -0.3292]

Current fitness CF(1)= $f(Cp(1)) = 1 + 2Cp(1) - Cp(1)^2$

= [0.1976, -6.7368, -2.5061, 0.7942, 1.0755]

Note: $Cp(1)^2$ is obtained by squaring individual elements of Cp(1). As of now, we obtained current velocity, current position and current fitness.

Local best position (LBP)of each particle up to first iteration is just its current position.

LBP(1)=Cp(1)=[-0.3425, 3.9558, -1.128, -0.0981, 0.0385]

Local Best fitness of each particle up to iteration 1=current fitness of iteration 1

Local Best Fitness (LBF)

LBF(1)=CF(1)=[0.1976, -6.7368, -2.5061, 0.7942, 1.0755]

Global Best Fitness of iteration 1= Max (LBF(1));

GBF(1)=1.0755 → for 5th particle

Global Best Position of iteration 1

GBP(1)=Corresponding current position of 5th particle in cp(1)

=0.0385

Velocity of iteration 2

Velocity for next iteration

$$V_{t+1} = WV_t + C_1 rand(0,1)(LBP(i) - Cp(i)) + C_2 rand(0,1)(GBP(i) - Cp(i))$$

We have from iteration 1

V(1)=[0.0319, 0.3185, 0.03331, 0.2677, -0.3292]

For 1st particle: r_1=0.4657 ,r_2=0.5319, CP(1)=-0.3425, LBP(1)=-0.3425 and GBP(1)=0.0385

So, for the iteration 2, for the particle 1st: $V_2 = 0.7V(1) + 0.2 * rand(0,1)\big(LBP(i) - Cp(i)\big) + 0.6 * rand(0,1)(GBP(i) - Cp(i))$ =0.1439

Thus we have for iteration 2

V(2)=[0.1439, -1.7008, 0.8136, 0.2503, -0.2304]

Current position and current fitness

Current position for next iteration

$$Cp(i+1) = cp(i) + V(i+1)$$

WKT,

CP(1)=[-0.3425, 3.9558, -1.1228, -0.0981, 0.0385] & V(2)=[0.1439, -1.7008, 0.8136, 0.2503, -0.2304]

Hence, CP(2)=[-0.1986, 2.2550, -0.3092, 0.1522, -0.1919]

Current fitness for next iteration

CF(i)= $f\big(Cp(i)\big) = 1 + 2Cp(i) - Cp(i)^2$

Hence, CF(2)=[0.5634, 0.4250, 0.2860, 1.2812, 0.5794]

We know that Local Best Fitness is LBF(1)=[0.1976, -6.7368, -2.5061, 0.7942, 1.0755]

Hence,

LBF(2)=Max[CF(2), LBF(1)] = [0.5634,0.4250, 0.2860, 1.2812, 1.0755]

Local Best & Global Best

We have for iteration 2:

CP(2)=[-0.1986, 2.2550, -0.3092, 0.1522, -0.1919] and LBF(2)= [0.5634,0.4250, 0.2860, 1.2812, 1.0755]

Hence Global Best Fitness in iteration 2,

GBF(2)= Max(LBF(2))=1.2812

So, Global Best Position in iteration 2, GBP(2)= 0.1522(4th particle position in CP(2))

Local Best Position of each particle in iteration 2

CP(1)=[-0.3425, 3.9558, -1.1228, -0.0981, 0.0385] and LBF(1)=[0.1976, 0.4250, 0.2860, 1.2816, 0.5794]

So, LBP(2)= position w.r.t. LBF(2)=[-0.1976, 2.2550, -0.3092, 0.1522, 0.0385]

Current position is best for first 4 particle, but not for 5th last one is better

Summary: Iteration 1 & 2

Iteration	V(i) & CP (i)	CF(i) & LBF (i)	GBF(i)	LBP(i) & GBP(i)
1	V(1)=[0.0319, 0.3185, 0.03331, 0.2677, -0.3292] CP(1)=[-0.3425, 3.9558, -1.1228, -0.0981, 0.0385]	CF(1)=[0.1976, -6.7368, -2.5061, 0.7942, 1.0755] LBF(1)=[0.1976, -6.7368, -2.5061, 0.7942, 1.0755]	GBF(1) =1.0755	LBP(1)=[-0.3425, 3.9558, -1.1228, -0.0981, 0.0385] GBF(1)=0.0385
2	V(2)=[0.1439, -1.7008, 0.8136, 0.2503, -0.2304]	CF(2)= [0.5634,0.4250, 0.2860, 1.2812, 0.5794]	GBF(2) =1.2812	LBP(2)=[-0.1986, 2.2550, -0.3092, 0.1522, 0.0385] GBP(2)=0.1522

	CP(2)=[-0.1986, 2.2550, -0.3092, 0.1522, -0.1919]	LBF(2)= [0.5634,0.4250, 0.2860, 1.2812, 1.0755]		

Summary: Iteration 3 & 4

3	V(3)=[0.02127, -2.2232, 0.8001, 0.1752, -0.1120] CP(3)=[0.0141, 0.0318, 0.4909, 0.3274, -0.2944]	CF(3)=[1.0279,1.0625,1.7410, 1.5464, 0.3246] LBF(3)=[1.0279, 1.0625, 1.7410, 1.5464, 1.0755]	GBF(3)=1.7410	LBP(3)=[0.0141, 0.0318, 0.4909, 0.3274, 0.0385] GBP(3)=0.4909
4	V(4)=[0.3011, -1.3308, 0.5601, 0.1980, 0.0420] CP(4)=[0.3152, -1.2990, 1.0510, 0.5254, -0.2523]	CF(4)=[1.5312, -3.2861, 1.9974, 1.7740, 0.4317] LBF(4)=[1.5312, 1.0625, 1.9974, 1.7740, 1.0755]	GBF(4)=1.9974 (Best fitness)	LBP(4)=[0.3152, 0.0318, 1.0510, 0.5254, 0.0385] GBP(4)=1.0510 (Best position)

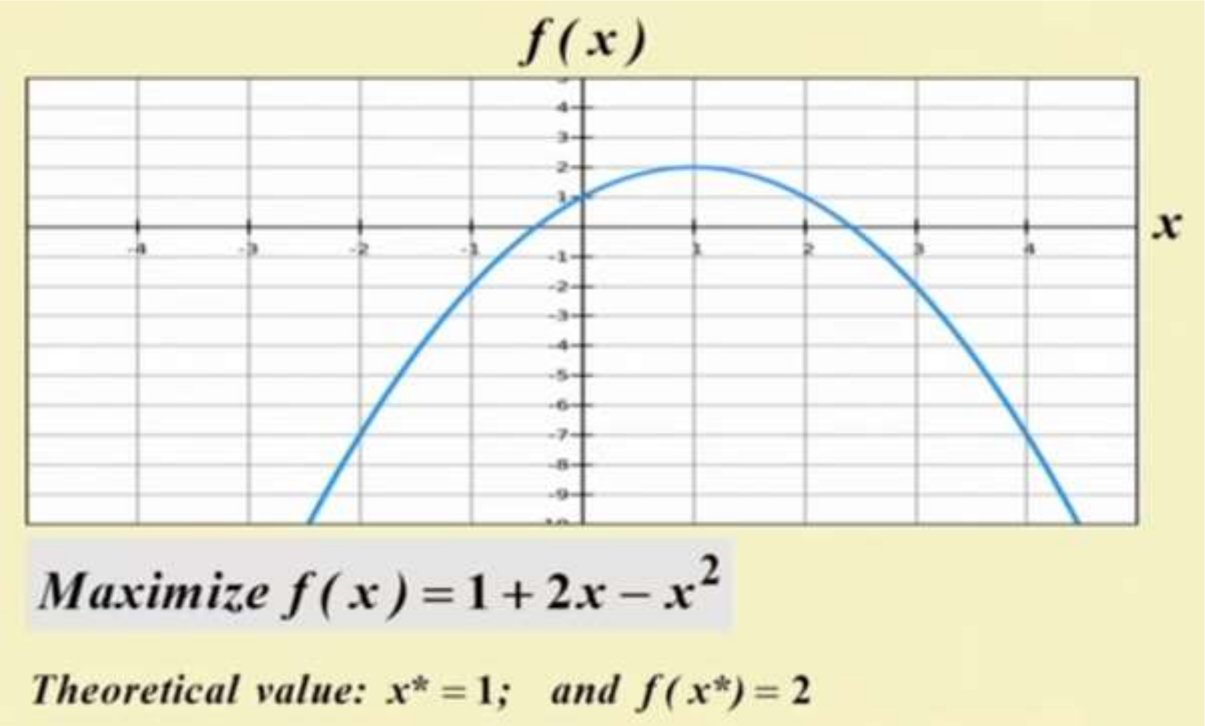
$$V_{t+1} = WV_t + C_1 rand(0,1)(pbest - X_t) + C_2 rand(0,1)(gbest - X_t)$$

$$Cp(i+1) = cp(i) + V(i+1)$$

$$LBF(i+1) = Max[CF(i+1), LBF(i)]$$

$$GBF(i) = Max[LBF(i)]$$

Final solution:



From iteration 4, we have, Global Best Position GBP(4)=1.0510 & Global Best Fitness GBF(4)=1.9974 Hence the final solution obtained as x*=1.0510 and f(x*)=1.9974.

Numerical Example 2 – Robust Regression with Particle Swarm Optimisation

[Reference: Enrico Schumann. Robust Regression with Particle Swarm Optimisation. https://cran.r-project.org/web/packages/NMOF/vignettes/PSlms.pdf]

```
#R code for – Robust Regression with Particle Swarm Optimisation
install.packages("NMOF")
install.packages("MASS")
library("NMOF")
library("MASS")
set.seed(11223344)
createData <- function(n, p, constant = TRUE,
                       sigma = 2, oFrac = 0.1) {
  X <- array(rnorm(n * p), dim = c(n, p))
  if (constant)
    X[, 1L] <- 1L
  b <- rnorm(p)
  y <- X %*% b + rnorm(n)*0.5
  nO <- ceiling(oFrac*n)
  when <- sample.int(n, nO)
  X[when, -1L] <- X[when, -1L] + rnorm(nO, sd = sigma)
  list(X = X, y = y, outliers = when)
}
n <- 100L ## number of observations
p <- 10L ## number of regressors
constant <- TRUE; sigma <- 5; oFrac <- 0.1
```

```
h <- 75L ## ... or use something like floor((n+1)/2)
aux <- createData(n, p, constant, sigma, oFrac)
X <- aux$X; y <- aux$y
Data <- list(y = as.vector(y), X = X, h = h)
plot(Data)
plot(X,y)
plot(y, type="l")
par(bty = "n", las = 1, tck = 0.01, mar = c(4,4,1,1))
plot(X[ ,2L], type = "h", ylab = "X values", xlab = "observation")
lines(aux$outliers, X[aux$outliers ,2L], type = "p", pch = 21,
      col = "blue", bg = "blue")
OF <- function(param, Data) {
  X <- Data$X; y <- Data$y
  aux <- y - X %*% param
  aux <- aux * aux
  aux <- apply(aux, 2L, sort, partial = Data$h)
  colSums(aux[1:Data$h, ]) ## LTS
}
popsize <- 100L; generations <- 500L
ps <- list(min = rep(-10,p),
           max = rep( 10,p),
           c1 = 0.9,
           c2 = 0.9,
           iner = 0.9,
           initV = 1,
           nP = popsize,
           nG = generations,
           maxV = 5,
           loopOF = FALSE,
           printBar = FALSE,
           printDetail = FALSE)
system.time(solPS <- PSopt(OF = OF, algo = ps, Data = Data))
solPS <- PSopt(OF = OF, algo = ps, Data = Data)
solPS
```

Suggested Readings:

Dai, H.-P.; Chen, D.-D.; Zheng, Z.-S. Effects of Random Values for Particle Swarm Optimization Algorithm. *Algorithms* 2018, *11*, 23. https://www.mdpi.com/1999-4893/11/2/23

Enrico Schumann. Robust Regression with Particle Swarm Optimisation. https://cran.r-project.org/web/packages/NMOF/vignettes/PSlms.pdf]

Gilli, M., D. Maringer and E. Schumann. (2011). *Numerical Methods and Optimization in Finance*. Elsevier.

J. Kennedy, The particle swarm: social adaptation of knowledge, IEEE International Conference on Evolutionary Computation, 1997Indianapolis, IN. https://ieeexplore.ieee.org/document/592326

Manfred Gilli, Dietmar Maringer, and Enrico Schumann. Numerical Methods and Optimization in Finance. Elsevier/Academic Press, 2011. URL http://enricoschumann.net/NMOF

Mohanty, P. (2018). NTPL online certification course on selected topics on decision modelling, Particle Swarm Optimization, IIT Khargapur. https://www.youtube.com/watch?v=uwXFnzWaCY0]

Soumya D. Mohanty (2012). Particle Swarm Optimization and regression analysis – I, Astronomical Review, 7:2, 29-35, DOI: 10.1080/21672857.2012.11519700

.

Artificial Neural Network approach for Time Series Forecasting

[1]Mrinmoy Ray, [1]K N Singh, [1]Kanchan Sinha and [2]Santosha Rathod
[1]ICAR-IASRI, New Delhi
[2]ICAR-IIRR, Hyderabad

Introduction:

An artificial neural network (ANN), otherwise called neural network (NN), is a computational device that is inspired by the structure as well as functional aspects of biological neural networks the human brain especially. A neural network made out of various interconnected simple processing elements called neurons or nodes. Each node receives an input signal which is the aggregate ‘‘information’’ from other nodes or external stimuli, processes it locally through an activation or transfer function and produces a transformed output signal to other nodes or external outputs. This information processing characteristic makes ANNs an effective computational device and able to learn from examples and then to generalize to examples never before seen.

A Time series (TS) is an ordered sequence of observations of a variable at equally spaced time intervals (monthly price data of a commodity, yearly crop yield and daily temperature data etc.). Time series forecasting is the utilization of a statistical model to predict future values based on previously observed values. The most widely used technique for forecasting time-series data is the Box Jenkins’ Autoregressive integrated moving average (ARIMA) methodology. ARIMA model is appropriate if the time series under study is linear. In any case, they might be absolutely inappropriate if the time series under investigation is nonlinear in nature. There are several nonlinear time series model to deal with nonlinear time series data for instance,, bilinear model, Threshold Autoregressive (TAR) model, Generalized Autoregressive Conditional Heteroscedastic (GARCH) model. Truth be told, these nonlinear models are still limited in that an explicit relationship for the data series at hand has to be hypothesized with little knowledge of the underlying law. In fact, the formulation of a nonlinear model to a particular data set is a very troublesome since there are too many possible nonlinear patterns and a prespecified nonlinear model may not be general enough to capture all the important features. Artificial neural networks, which are nonlinear data-driven approaches as opposed to the above model-based nonlinear methods, are capable of performing nonlinear modeling without a priori knowledge about the relationships between input and output variables. In this way they are a more general and flexible modeling tool for forecasting. Thusly, in time series forecasting parlance, the ANN is a nonparametric nonlinear statistical model.

Overview of ANN architecture:

In general, an ANN can be partitioned into three sections, named layers, which are known as:

i) Input layers

These layers are responsible for receiving information (data), signals, features, or measurements from the external environment. These inputs (samples or patterns) are usually normalized within the limit values produced by activation functions. This normalization results in better numerical precision for the mathematical operations performed by the network.

ii) **Hidden, intermediate, or invisible layers**

These layers are composed of neurons which are responsible for extracting patterns associated with the process or system being analyzed. These layers perform most of the internal processing from a network.

iii) Output layers

These layers are also composed of neurons, and thus are responsible for producing and presenting the final network outputs, which result from the processing performed by the neurons in the previous layers.

The main architectures of artificial neural networks, considering the neuron disposition, as well as how they are interconnected and how its layers are composed can be divided as follows:

a) Single-layer feedforward network

b) Multilayer feedforward networks

c) Recurrent networks

d) Mesh networks.

For time series forecasting the multilayer feedforward networks are used which is given below

Multilayer feedforward networks

Figure 1 shows a feedforward network with multiple layers composed of one input layer with n 3sample signals, two hidden neural layers consisting of n1 and n2 neurons respectively, and, finally, one output neural layer composed of m neurons representing the respective output values of the problem being analyzed.

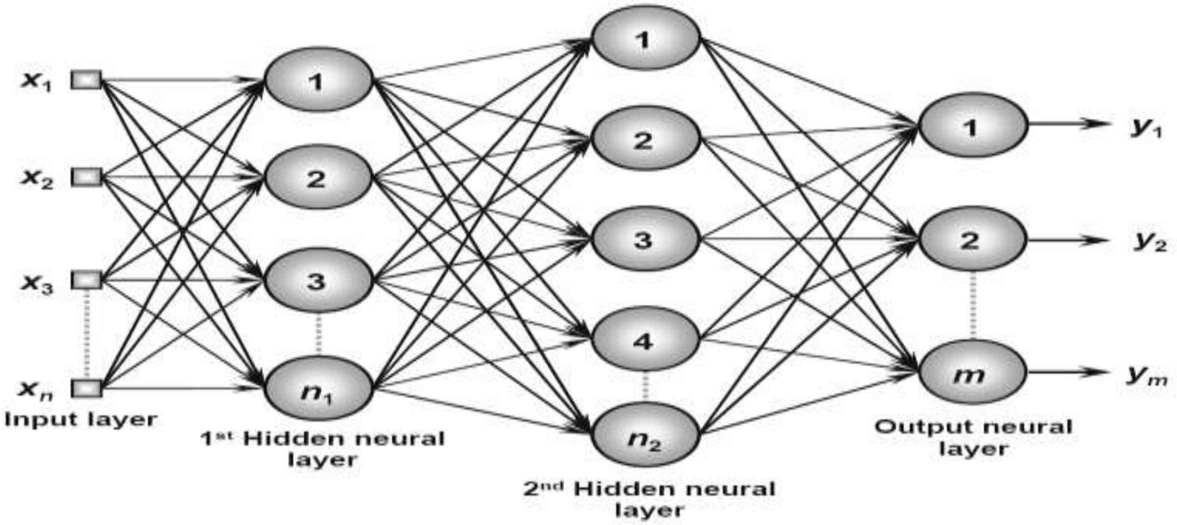


Fig 1: Architecture of Multilayer feedforward networks

ANN approach to time series forecasting:

In the domain of time series analysis, the inputs are typically the past observations series and the output is the future value. The ANN performs the following nonlinear function mapping between the input and output

$$y_t = f(y_{t-1} + y_{t-2}, ..., y_{t-p}, w) + \varepsilon_t$$

where, w is a vector of all parameters and f is a function of network structure and connection weights. Therefore, the neural network resembles a nonlinear autoregressive model.

Single hidden layer multilayer feed forward network is the most popular for time series modeling and forecasting. This model is characterized by a network of three layers of simple processing units. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer.

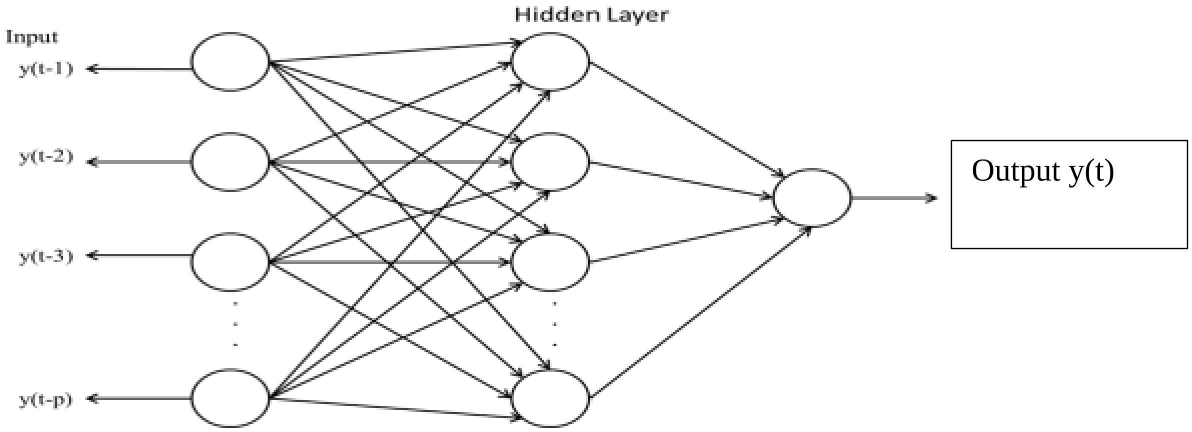


Fig 2: Architecture of ANN for time series forecasting

The relationship between the output (y_t) and the inputs (y_{t-1}, $y_{t-2}, \ldots, y_{t-p}$) can be mathematically represented as follows:

$$y_t = f\left(\sum_{j=0}^{q} \omega_j g\left(\sum_{i=0}^{p} \omega_{ij} y_{t-i}\right)\right)$$

where, $\omega_j(j = 0,1,2, \ldots\ldots, q)$ and $\omega_{ij}(i = 0,1,2, \ldots\ldots, p,\ j = 0,1,2, \ldots\ldots, q)$are the model parameters often called the connection weights, p is the number of input nodes and q is the number of hidden nodes, g and f denote the activation function at hidden and output layer respectively. Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity. Most commonly used activation functions are as follows-

Activation function	**Equation**
Identity	x
Sigmoid	$\frac{1}{1+e^{-x}}$
TanH	$\tanh(x) = \frac{2}{1+e^{-2x}} - 1$
ArcTan	$\tan^{-1}(x)$
Sinusoid	$\sin(x)$
Gaussian	e^{-x^2}

For time series forecasting sigmoid activation function is employed in hidden layer and identity activation function is employed in the output layer.

The selection of appropriate number of hidden nodes as well as optimum number of lagged observation *p* for input vector is important in ANN modeling for determination of the autocorrelation structure present in a time series. Though there are no established theories available for the selection of *p* and *q*, hence experiments are often conducted for the determination of the optimal values of *p* and *q*. The connection weights of ANNs are determined by learning method. There are three common learning algorithms for ANN –

1) Supervised Learning

The supervised learning strategy consists of having available the desired outputs for a given set of input signals; in other words, each training sample is composed of the input signals and their corresponding outputs. Henceforth, it requires a table with input/output data, also called attribute/value table, which represents the process and its behavior.

2) Unsupervised Learning

Different from supervised learning, the application of algorithm based on unsupervised learning does not require any knowledge of the respective desired outputs. Thus, the network needs to organize itself when there are existing particularities between the elements that compose the entire sample set, identifying subsets (or clusters) presenting similarities. The learning algorithm adjusts the synaptic weights and thresholds of the network in order to reflect these clusters within the network .itself.

3) Reinforcement Learning

It is the hybrid of supervised and unsupervised learning.

For time series forecasting supervised learning approach is utilized. Gradient decent back propagation algorithm is one of the popular approach of supervised learning.

Gradient decent back propagation algorithm

The objective of training is to minimize the error function that measures the misfit between the predicted value and the actual value. The error function which is widely used is mean squared error which can be written as:

$$E = \frac{1}{N}\sum_{n=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{n=1}^{N}\left\{y_t - f\left(\sum_{j=0}^{q}\omega_j g\left(\sum_{i=0}^{p}\omega_{ij} y_{t-i}\right)\right)\right\}^2$$

Where *N* is the total number of error terms. The parameters of the neural network are ω_j and $\omega_{i\,j}$ estimated by iteration. Initial connection weights are taken randomly from uniform distribution. In each iteration the connection weights changed by an amount $\Delta\omega_j$

$$\Delta\omega_j(t) = -\eta\frac{\partial E}{\partial \omega_j} + \delta\Delta\omega_j(t-1)$$

where, η is the learning rate and $\frac{\partial E}{\partial \omega_j}$ is the partial derivative of the function E with respect to the weight ω_j. δ is the momentum rate. The $\frac{\partial E}{\partial \omega_j}$ can be represented as follows-

$$\frac{\partial E}{\partial w_j} = -e_j(n) \times f'(x) \times y_j(n)$$

where $e_j(n)$ is the residual at nth iteration

$f'(x)$= derivative of the activation function in the output layer. As in time series forecasting the activation function in the output layer is identity function hence $f'(x)$=1. $y_j(n)$ is the desired output. Now connection weights in from input to hidden nodes changed by an amount $\Delta\omega_{ij}$

$$\Delta\omega_{ij}(t) = -\eta\frac{\partial E}{\partial \omega_{ij}} + \delta\Delta\omega_{i\,j}(t-1)$$

where

$$\frac{\partial E}{\partial w_{ij}} = g'(x) \times \sum_{j=0}^{q} e_j(n) * w_j(n)$$

where $g'(x)$ is the activation function in the hidden layer. For sigmoid activation function

$$g'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

Learning rate is user defined parameter known as tuning parameter of neural network which determine how slow or fast the optimal weight is obtained. The learning rate must be set small enough to avoid divergence. The momentum term prevents the learning process from setting in a local minimum. Though there are no established theories available for the selection of learning rate and momentum, hence experiments are often conducted for the determination of the learning rate and momentum.

Step by Step Modeling Procedure:

1. Testing of Nonlinearity:

As ANNs is suitable for nonlinear time series forecasting. Hence, prior to application of ANN the nonlinearity should be check. There are several tests for checking nonlinearity. BDS (Brock-Dechert-Scheinkman) test is of the popular approach for checking nonlinearity. This test utilizes the concept of spatial correlation from chaos theory. The computational procedure is given as follows

i) Let the considered time series is

$$\{x_i\} = [x_1, x_2, x_3, ..., x_N]$$

ii) The next step is to specify a value of m (embedding dimension), embed the time series into m dimensional vectors, by taking each m successive points in the series. This transforms the series of scalars into a series of vectors with overlapping entries

$$x_1^m = (x_1, x_2, ..., x_m)$$
$$x_2^m = (x_2, x_3, ..., x_{m+1})$$
$$.$$
$$.$$
$$.$$
$$x_{N-m}^m = (x_{N-m}, x_{N-m+1}, ..., x_N)$$

iii) In the third step correlation integral is computed, which measures the spatial correlation among the points, by adding the number of pairs of points (*i, j*), where 1≤ *i* ≤ N and 1≤ *j*≤N , in the m-dimensional space which are “close” in the sense that the points are within a radius or tolerance ε of each other.

$$C_{\varepsilon,m} = \frac{1}{N_m(N_m - 1)} \sum_{i \neq j} I_{i,j;\varepsilon}$$

where $I_{i,j;\varepsilon}$= 1 if $\left\| x_i^m - x_j^m \right\| \leq \varepsilon$

= 0 otherwise

iv) If the time series is i.i.d. then $C_{\varepsilon,m} \approx [C_{\varepsilon,1}]^m$

v) The BDS test statistics is as follows

$$BDS_{\varepsilon,m} = \frac{\sqrt{N}[C_{\varepsilon,m} - (C_{\varepsilon,1})^m]}{\sqrt{V_{\varepsilon,m}}}$$

where $V_{\varepsilon,m} = 4[K^m + 2\sum_{j=1}^{m-1} K^{m-j} C_\varepsilon{}^{2j} + (m-1)^2 C_\varepsilon{}^{2m} - m^2 K C_\varepsilon{}^{2m-2}]$

$$K = K_\varepsilon = \frac{6}{N_m(N_m - 1)(N_m - 2)} \sum_{i<j<N} h_{i,j,N;\varepsilon}$$

$$h_{i,j,N;\varepsilon} = \frac{[I_{i,j;\varepsilon} I_{j,N;\varepsilon} + I_{i,N;\varepsilon} I_{N,j;\varepsilon} + I_{j,i;\varepsilon} I_{i,N;\varepsilon}]}{3}$$

The choice of m and ε depends on number of data. The null hypothesis is data are independently and identically distributed (i.i.d) against the alternative hypothesis the data are not i.i.d.; this implies that the time series is non-linearly dependent. BDS test is a two-tailed test; the null hypothesis should be rejected if the BDS test statistic is greater than or less than the critical values.

2. Division of the data:

Data is divided into training and test sets. The training sample is used for ANN for model development and the test sample is utilized to evaluate the forecasting performance. Sometimes a third one called the validation sample is also utilized to avoid the overfitting problem or to determine the stopping point of the training process. It is common to use one test set for both validation and testing purposes particularly for small data sets. The literature suggests little guidance in selecting the training and testing sets. Most commonly used rule are 90% vs. 10%, 80% vs. 20% or 70% vs. 30%, etc.

3. Data Normalization:

Nonlinear activation functions such as the sigmoid function typically have the squashing role in restricting the possible output from a node to, typically, (0, 1). Hence, data normalization is done prior to training process begins.

Normalization procedure

Linear transformation to [0,1]: X_n =(X_0-X_{min})/ (X_{max}-X_{min})

Statistical normalization: X_n =(X_0-mean(X))/var(X)

simple normalization: X_n =X_0/ X_{max}

4. Selection of appropriate number of hidden nodes as well as optimum number of lagged:

There are no established theories available for the selection of *p* and *q*, hence experiments are often conducted for the determination of the optimal values of *p* and *q*.

5. Estimation of connection weights:

Estimation of connection weights are determined by learning algorithm. For time series forecasting most commonly used learning approach is gradient decent back propagation algorithm.

6. Evaluating forecasting Performance

Forecasting performance can be computed by several approaches. Some of the approaches are given below-

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| / y_t \times 100$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}$$

where n is the total number of forecast values. y_t is the actual value at period t and $\hat{y}_t$ is the corresponding forecast value. The model with less MAPE/MSE/RMSE is preferred for forecasting purposes.

Limitations of ANN for time series forecasting:

i) ANNs are nonlinear time series model hence, for linear time series data the approach may not be better than linear statistical model.

ii) ANNs are black-box methods. There is no exact form to describe and analyze the relationship between inputs and outputs. This causes troublesome for interpretation of results. In addition, no formal statistical test is available.

iii) ANNs are subjected to have overfitting problems owing to its large number of parameters.

iv) There are no established theories available for the selection of p and q, hence experiments are often conducted for the determination of the optimal values of p and q which is tedious.

v) ANNs usually require more data for time series forecasting.

References:

Anjoy, P., Paul, R. K., Sinha, K., Paul, A. K. and Ray, M. (2017). A hybrid wavelet based neural networks model for predicting monthly WPI of pulses in India. *Indian Journal of Agricultural Sciences*. **87 (6)**, 834-839.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2009), *Time Series Analysis: Forecasting and Control (3rd ed.), San Francisco: Holden-Day.*

Broock, W., Scheinkman, J. A., Dechert, W. D. and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Review,* **15,** 197–235.

Jha, G. K., and Sinha, K.2014.Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications* 24 (3): 563-571.

Makridakis, S., Wheelwright, S.C. and Hyndman, R. J. (1998).*Forecasting: Methods and Applications (3rd ed.),* Chichester: Wiley.

Mukherjee, A., Rakshit, S., Nag, A., Ray, M**.,** Kharbikar, H. L., Kumari, S., Sarkar, S., Paul, S., Roy, S., Maity, A., Meena, V. S. and Burman, R. R. (2016). Climate Change Risk Perception, Adaptation and Mitigation Strategy: An Extension Outlook in Mountain Himalaya. In: Jaideep Kumar Bisht, Vijay Singh Meena, Pankaj Kumar Mishra and Arunava Pattanayak Edition. Conservation Agriculture (pp. 257-292). Singapore. Springer Singapore.

Ray, M., Rai, A., Ramasubramanian, V. and Singh, K. N. (2016). ARIMA-WNN hybrid model for forecasting wheat yield time series data. *Journal of the Indian Society of Agricultural Statistics*, **70(1)**, 63-70.

Ray, M., Rai, A., Singh, K. N., Ramasubramanian, V. and Kumar, A. (2017). Technology forecasting using time series intervention based trend impact analysis for wheat yield scenario in India. *Technological Forecasting & Social Change*, **118**, 128–133.

Remus, W. and O'Connor, M.(2001). *Neural Networks for Time-Series Forecasting, New york, Springer.*

Zhang, G., Patuwo, B. E. and Hu, M. Y. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14: 35-62.

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**, 159–175.

Support Vector Machine: A Non-Linear Machine Learning Technique

Amit Saha[1], K. N. Singh[2], Mrinmoy Ray[2] and Santosha Rathod[3]
[1]Central Silk Board, Ministry of Textiles, Government of India
[2]ICAR-IASRI, New Delhi
[3]ICAR-IIRR, Hyderabad
amits.csb@gov.in

1. Introduction:

Machine learning is a technique which allows the machine to learn by itself. Support Vector Machine (SVM) is one of the eminent supervised machine learning technique which was developed by Cortex and Vapnik (1995) for binary classification problems. In binary classification, the goal of the SVM is to find out a hyperplane that best separates a dataset into two classes. After two years of SVM's invention, support vector regression (SVR) based on similar principles as SVM classification was developed by Vapnik *et al*. (1997) to deal with the regression problems. Being a non-parametric method, SVR does not depend on assumptions like linear regression. Another benefit of using SVR is that it permits the construction of non-linear model. So, SVM is not only popular for the classification but also for its modelling and prediction ability. The performance of SVM is based upon proper selection of kernel. There are different types of kernel which can be used for the classification and prediction purposes. Since the last decade, the application of SVM has been extended to time series modelling and forecasting in various areas such as power load forecasting (Niu *et al*., 2010), rainfall forecasting (Ortiz-Garcia et al., 2014), wind power forecasting (De Giorgi *et al*., 2014) and agricultural forecasting (Kumar and Prajneshu, 2015).

2. Support Vector Machine (SVM) in time series:

Application of SVM in time series is generally utilized when the series shows non stationarity and non-linearity process. A tremendous advantage of SVM is that it is not model dependent as well as independent of stationarity and linearity. However, it may be computationally expensive during the training. The training of the data driven prediction process SVM is done by a function which is estimated utilizing the observed data. Let, a time series $y(t)$ which takes the data at time t $\{t = 0,1,2,3, \dots, N\}$.

Now, the prediction function for linear regression is defined as:

$$f(y) = (w.y) + c \tag{1}$$

Whereas, for non linear regression, it will be:

$$f(y) = \big(w.\emptyset(y)\big) + c \tag{2}$$

Where, w dentoes the weights, c represents threshold value and $\emptyset(y)$ is known as kernel function.

If the observed data is linear, then equation (1) will be used. But, for non-linear data, the mapping of $y(t)$ is done to the higher dimension feature space through some function which is denoted as $\emptyset(y)$ and eventually it is transformed into the linear process. Afer that, a linear regression will carry out in that feature space.

The first and foremost objective is to find out the value of w and c which will be optimal. In SVM, there are two things viz., flatness of weights and error after the estimation which are to be minimized. The flatness of the weights is denoted by $\|w\|^2$ which is the eucledian norm. Firstly, one has to concentrate on minimization the $\|w\|^2$. Second important thing is the minimization of the error. This is also called as empirical risk. However, the overall aim is to minimize the regularized risk which is sum of empirical risk and the half of the product of the flatness of weight and a constant term which is known as regularized constant. The regularized risk can be written as-

$$R_{reg}(f) = R_{emp}(f) + \tfrac{\tau}{2}\|w\|^2 \tag{3}$$

Where, $R_{reg}(f)$ is the regularized risk, $R_{emp}(f)$ denotes the empirical risk, τ is as constant which is called as regularized constant/capacity control term and $\|w\|^2$ is the flatness of weights.

The regularization constant has a significant impact on a better fitting of the data and it can also be useful for the minimization of bad generalization effects. In the other words, this constant deals with the problem of over-fitting. The overfitting of the data can be redued by the proper selection of this constant value. The empirical risk can be defined as:-

$$R_{emp}(f) = \tfrac{1}{N}\textstyle\sum_{i=0}^{N-1} L\big(y(i), \alpha(i), f(y(i), w)\big) \tag{4}$$

Where, $\alpha(i)$ denotes the truth data of predicted value, $L(.)$ is known as loss function and i represents the index to the time series.

There are various types of loss function in literature. But, two functions viz., vapnik loss function and quadratic loss function are most popular and they are generally used. The quadratic programming problem has been made to minimize the regularised risk which is-

$$\text{Minimize, } \frac{1}{2}\|w\|^2 + D\sum_{i=1}^{n} L\big(\alpha(i), f(y(i), w)\big) \qquad (5)$$

Where,

$$L\big(\alpha(i), f(y(i), w)\big) = |\alpha(i) - f(y(i), w)| - \in \text{ if } |\alpha(i) - f(y(i), w)| \geq \in$$

$$= 0; \text{ otherwise.}$$

Where, D is a constant which equals to the summation normalization factor and $\in$ represents the size of the tube.

The computation of $\in$ and D is done empirically because they are user defined. On has to choose proper value of D and $\in$. Now, dual optimization problem is formed using the lagrange multiplier which can be written as:

$$\text{Maximize, } -\frac{1}{2}\sum_{i,j=1}^{N}(\beta_i - \beta_i{}^*)\left(\beta_j - \beta_j{}^*\right)\langle y(i), y(j)\rangle - \in \sum_{i=1}^{N}(\beta_i - \beta_i{}^*) + \sum_{i=1}^{N}\alpha(i)\,(\beta_i - \beta_i{}^*) \qquad (6)$$

Subject to, $\sum_{i-1}^{N}(\beta_i - \beta_i{}^*) = 0$; $\beta_i, \beta_i{}^* \in [0, D]$

The function $f(x)$ is defined as;

$$f(x) = \sum_{i=1}^{N}(\beta_i - \beta_i{}^*)\,\langle y, y(i)\rangle + C$$

(7)

KKT conditions are used to get the solution of the weights.

The significance of kernel function in non-linear support vector machine (NLSVR) is very much imporatnt for mapping the data $y(i)$ into higher dimension feature space $\emptyset\big(\,y(i)\big)$ in which the data becomes linear. Generally notation for kernel function is given as;

$$k(y, y') = \langle\emptyset(y), \emptyset(\,y')\rangle; \qquad (8)$$

There are many methods in literature to solve the quadartic programming. However, the most used method is sequential minimization optimization (SMO) algorithm.

3. Kernel function

SVM is a learning algorithm which is based on kernel. There are different types of kernel which can be used for the classification and prediction purpose. However, there is no such rule to make inference on which kernel should one use. All the kernels are used separately for the given datasets and whichever gives the better result, one should choose that one. Various types Kernel are listed below:

1. Non linear
2. Linear
3. Polynomial
4. Radial basis function: a) Gaussian Radial basis function b) Laplace Radial basis function.
5. Sigmoid kernel
6. Hyperbolic tangent kernel
7. Anova radial basis kernel
8. Multi-layer perceptron
9. Linear spline kernel.

Kernel function are used for the transformation of the given data into the required form. Kernel function is actually a mathematical function. RBF is mostly used kernel function. Some kernel functions are described in the following:

Polynomial kernel equation: Polynomial kernel is generally used in the image processing. It is useful for nonlinear modelling. This kernel function is very simple yet efficient method.

$$k(x,y) = (x.y+1)^p \text{ ; } p\text{= degree of polynomial} \qquad (9)$$

Gaussian kernel function:

$$k(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \qquad (10)$$

Or $k(x, y) = exp(-\alpha \|x - y\|^2)$, Where, shape of hyperplane is controlled by σ.

Sigmoid kernel function:

Sigmoid function is used as the proxy of artificial neural network.

$$k(x, y) = tanh(\theta x^T . y + a) \tag{11}$$

Linear kernel function:

Sometimes, linear kernel gives better results as compared to complex and nonlinear kernels. Linear

classifier can be used to test the non-linearity of the datasets.

$$k(x, y) = x . y \tag{12}$$

4. Advantages of SVM:

1. It gives global optimum.
2. Training of SVM is comparatively easier than other machine learning techniques.
3. Well scaling for data with high dimensionality.
4. It can give a good prediction.
5. It is based on statistical learning theory.
6. Work on structural risk minimization.
7. Risk of overfitting problem may overcome by SVM.
8. It has good generalization property.
9. It is useful when there is no prior information about the data.
10. It also work on unstructured data.

5. Illustration:

Data Description:

Time series data on Cotton Production (Million Bales) of India from 1950 to 2016 were taken from the Ministry of Agriculture & Farmers Welfare, Government of India. The data from

1950-2011 have been utilized for model building purpose and the data from 2012 to 2016 were used to predict the cotton production for the validation purpose.

Support Vector Machine:

The most important part in SVM technique is the selection of parameters and kernel which have to be selected with utmost care to improve the performance of the model in order to get better accuracy in forecasting. The best parameters and kernel have been selected using “e1701” package (David, 2017) in R software.

The time series plot of cotton production is illustrated in Fig. 1. It can be seen from Table 1 that the time series show a high value of coefficient variation which reprsents the presence of highly heterogenous characteristic of the series.

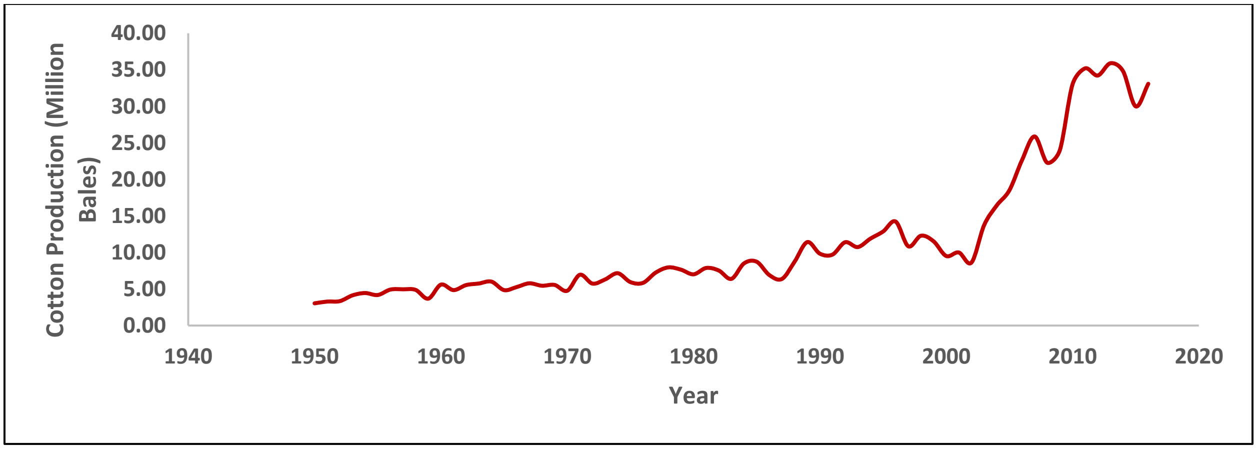


Fig. 1: Time Series Plot of Cotton Production

Table 1: Summary Statistics of Cotton Production

Statistic	Value	Statistic	Value
Minimum	3.04	Maximum	33.20
1st Quartile	5.54	Standard Deviation	6.81
Median	7.20	Skewness	2.05
Mean	9.60	Kurtosis	4.09
3rd Quartile	11.26	Coefficient of Variation	70.93

Table 2 displays the estimated best parameters of SVR after sufficient tuning of SVR model and these best parameters have been utilized to build the SVR model. It has been seen that the best SVM-kernel function is Radial basis function for SVR.

Table 2: Parameter estimation of SVR

Sampling method	10-fold cross validation
Epsilon (Best Parameter)	0.1
Cost (Best Parameter)	4
Gamma (Best Parameter)	1
Number of Support Vectors	39
SVM-Type	eps-regression
SVM-Kernel	Radial Basis Function

Fig. 2 shows the graphical representation of the performance of the models for Cotton Production series. Model performance in terms of MSE, MAE and MAPE has been shown in Table 3 and Table 4 for training and testing dataset respectively. Here, ARIMA (2, 2, 1) model has been fitted based on the lowest AIC values among various ARIMA models and the data of cotton production show the non-linearity pattern which is tested by Brock, Dechert and Scheinkman (BDS) test.

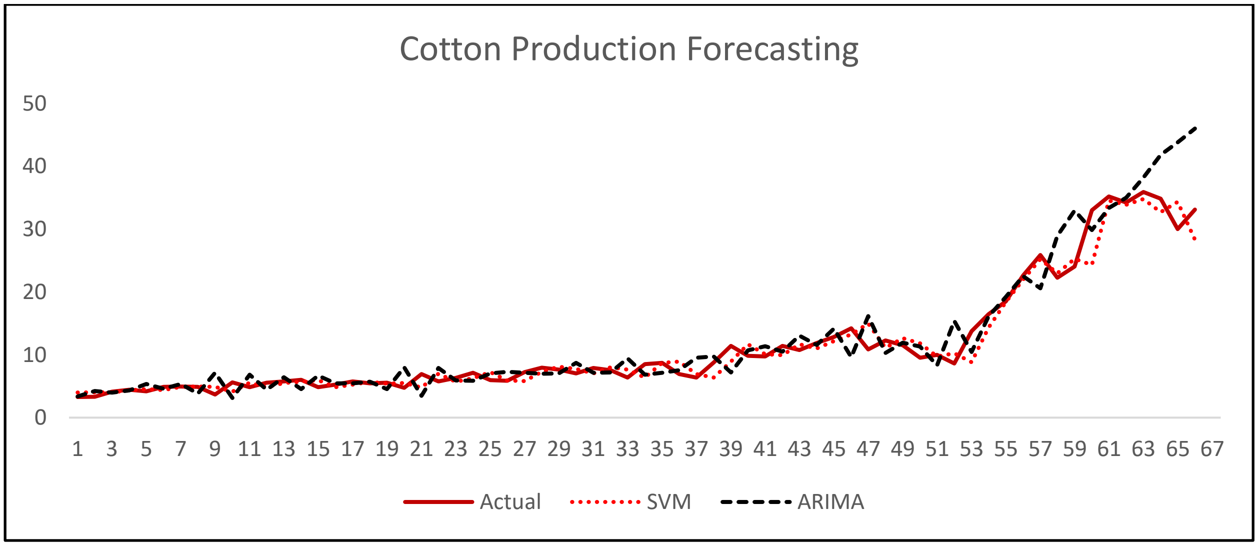


Fig. 2: Graphical representation of the performance of ARIMA and SVM models

Table 3: Model performance in training dataset using ARIMA and SVM

Model	MSE	MAE	MAPE
ARIMA	6.70	1.83	21.28
SVM	3.08	1.14	12.73

Table 4: Model performance in testing dataset using ARIMA and SVM

Model	MSE	MAE	MAPE
ARIMA	82.45	7.35	22.76
SVM	9.48	2.54	7.83

Table 5 displays the Out-of-Sample forecast values using ARIMA and SVM.

Table 5: Model performance in testing dataset using ARIMA and SVM

Year	Actual	ARIMA	SVM
2012	34.22	34.98	33.85
2013	35.9	38.21	34.78
2014	34.81	41.79	32.67
2015	30.0	43.82	34.33
2016	33.09	45.99	28.32

It has been seen from the Fig. 2 that the fitted graph of the SVM model is more close to the graph of original data as compare to ARIMA model both in training and forecasting. It is observed from Table 3 and Table 4 that the SVM has a lower MSE, MAE and MAPE compared to the ARIMA model in both training and testing dataset. It has also been seen from Table 5 that the forecasted values of the SVM are closer to the observed values compared to ARIMA. From the above results and discussion, it can be inferred that performance of the SVM model is better than the ARIMA model in terms of forecasting accuracy.

6. Conclusion

In reality, most of the time series data are non-linear in nature. In this study, the data of cotton production show non-stationary as well as non-linearity structure which were difficult to capture for the ARIMA models. However, SVM has shown its' tremendous performance due to the ability of capturing the non-linear pattern. Being a non-linear machine learning technique, SVM has well captured the heterogeneous trend of the given dataset. Based on the results, it can be inferred that SVM outperformed the ARIMA model. Therefore, it can be used in modeling and forecasting of time series to improve the forecasting accuracy in the presence of non-linear pattern.

7. Bibliography

Cortes, C. and Vapnik, V. (1995). Support-vector network. *Machine Learning*, 20, 1-25.

David, M. (2017). E1071: Misc Functions of the Department of Statistics. *Probability Theory Group R package version*, 1: 6–8.

De Giorgi, M.G., Campilongo, S., Ficarella, A. and Congedo, P.M. (2014). Comparison between wind power rediction models based on wavelet decomposition with least-squares support vector machine (LS-SVM) and artificial neural network (ANN). *Energies,* 7:5251-5272.

Kumar, T.L.M. and Prajneshu (2015). Development of Hybrid Models for Forecasting Time-Series Data Using Nonlinear SVR Enhanced by PSO. *Journal of Statistical Theory and Practice*, 9(4), 699-711.

Niu, D., Wang, Y. and Wu, D.D. (2010). Power load forecasting using support vector machine and ant colony optimization. *Expert Syst Appl,* 37:2531–2539.

Ortiz-Garcia, E.G., Salcedo-Sanz, S. and Casanova-Mateom, C. (2014). Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmos Res*, 139:128–136.

Vapnik, V., Golowich, S., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing, In Mozer, M., Jordan, M and Petsche, T. (Eds). *Advances in Neural Information Processing Systems*, 9:281-287, Cambridge, MA, MIT Press.

Spatio-temporal Time Series Modelling and Forecasting

Santosha Rathod[1] and Mrinmoy Ray[2]
1-ICAR- IIRR, Hyderabad
2- ICAR- IASRI, New Delhi
santosha.rathod@icar.gov.in

Introduction:

Spatio-temporal time series are the observations which are recorded over both space and time by considering systematic dependencies across space and time. Spatio-temporal modeling manages the single variables recorded over a timeframe at various locations. The case of spatio-temporal data incorporates; Daily or hourly carbon emission data recorded from observatory at many location, daily river flow data recorded from many river basins, hourly daily or weekly record of many weather parameters over different locations, and traffic flow measurements taken from a set of loop detectors in an exceptionally visit premise are cases of spatial time series data. Utilization of the spatio-temporal time series modeling for the cover many areas and much of the original impetus for the area was driven by geo-statistics yet as of late the applications have been reached out to numerous areas viz., sociology, economics, environmental, ecological and agricultural sciences. Many literatures recommend that incorporation of both spatial and temporal information will enhance the demonstrating effectiveness of phenomenon under thought. In this way, it is sensible to model time and space scales at the same time to catch inherent vulnerability over a timeframe over various locations.

Because of computational difficulties and inaccessibility of simultaneous spatial and temporal information, no significant progress is accomplished in spatio-temporal time series modeling as contrast with univariate time series modeling. Spatio-temporal models are the models which considers concurrent information on both space and time of variables under consideration. In univariate time series we observe autocorrelation between the successive observations over a timeframe, to model these sorts of series, the Box-Jenkins autoregressive moving average (Box and Jenkins (1970)) model is most usually utilized model because of its prominent modeling building process. On other hand, the auto correlated spatio-temporal time series phenomenon can be modeled using the space time autoregressive moving average (STARMA) model. The autoregressive and moving average components of univariate time series lagged in both space and time is alluded as space time autoregressive moving average (STARMA) model.

1.1.2. STARMA Model

The space-time models explain the systematic dependencies over both space and time is modeled through the class of STARMA models was developed by Pfeifer and Deutsch (1980b). The autoregressive and moving average form of space time model represented by STARMA model are characterized by single variable *Z_i(t)*, observed at *N* fixed spatial locations (*i* = 1, 2,…, *N*) on *T* time periods (*t* = 1, 2, . . ., *T*). The *N* spatial locations can be a geographical location, country, state, *etc*. The spatial dependencies between N times series is incorporated through *N*N* spatial weight matrices. Analogous to univariate time series, ***Z(t)*** is expressed as a linear combination of past observations and errors. The STARMA model (Pfeifer and Deutsch, 1980a), denoted by $STARMA(p_{\lambda_1,\lambda_2,\dots,\lambda p}, q_{m_1,m_2,\dots,\mathrm{m}q})$ can be represented in the matrix equation as follows;

$$Z(t) = \sum_{K=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}\, W^l\, Z(t-k) - \sum_{K=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}\, W^l\, \varepsilon(t-k) + \varepsilon(t) \qquad \dots (1.1)$$

Where,

$\boldsymbol{z(t)} = [\boldsymbol{z_1}(t),\ \dots\dots, \boldsymbol{z_N}(t)]'$is a $N \times 1$ vector of observations at time $t = 1, \dots, T$,

p is the autoregressive order (AR) with respect to time,

q is the moving average order (MA) with respect to time,

λ_k is the spatial order of the k^{th} AR term,

m_k is the spatial order of the k^{th} MA term,

ϕ_{kl} is the AR parameter at temporal lag k and spatial lag *l* (scalar),

θ_{kl} is the MA parameter at temporal lag k and spatial lag *l* (scalar) and

$W^l$is the *N*N* spatial weight matrix with spatial order *l* with diagonal elements zero and non-diagonal elements is the relation between sites.

The spatial weight matrix $W^{(0)} = I_N$ *i.e.* Identity matrix and each row of W^l must add up to one. The random error vector $\varepsilon(t) = [\varepsilon_1(t), \varepsilon_2(t), \dots, \varepsilon_N(t)]'$ is normally distributed at time *t* with $E[\varepsilon(t)] = 0$, $E[\varepsilon(t)\varepsilon'(t+s)] = \begin{cases} G = \sigma^2 I_N \, is \; s = 0 \\ 0, \; otherwise \end{cases}$ and $E[\varepsilon(t)\varepsilon'(t+s)] = 0, for \; s > 0.$

There are two subclasses of the STARMA model, in equation (3) when *q*=0, only autoregressive terms remain and consequently the model progresses toward becoming space-time autoregressive or STAR model which is represented as follows;

$$\boldsymbol{Z(t)} = \sum_{K=1}^{p} \sum_{l=0}^{\lambda_k} \phi_{kl} \, W^l \, Z(t-k) + \varepsilon(t) \qquad \ldots(1.2)$$

When p becomes 0, only moving average terms remains and hence the model becomes space-time moving average or STMA model which is represented as follows;

$$\boldsymbol{Z(t)} = \varepsilon(t) - \sum_{K=1}^{q} \sum_{l=0}^{m_k} \theta_{kl} \, W^l \, \varepsilon(t-k) \qquad \ldots(1.3)$$

1.1.3. Spatial weight matrix

Building of spatial weight matrix plays a key role in STARMA modeling, the hierarchical ordering of neighbors of each site and the selection of an appropriate sequence of weighting matrices is a matter left to the model builder since more complex the weight matrix, more troublesome is to estimate the parameters of STARMA model. In the vast majority cases, the space pattern is assumed to be equal and regularly spaced to ease the model building. In the vast majority applications, the uniform spatial weight matrix is only a simplifying assumption since typically the sites are irregularly spaced. A weight can be picked in different ways, the least difficult of which is the binary scheme, if two areas shared a common border then we relegate a weight as one otherwise zero (Griffith (1996) and (2009)).

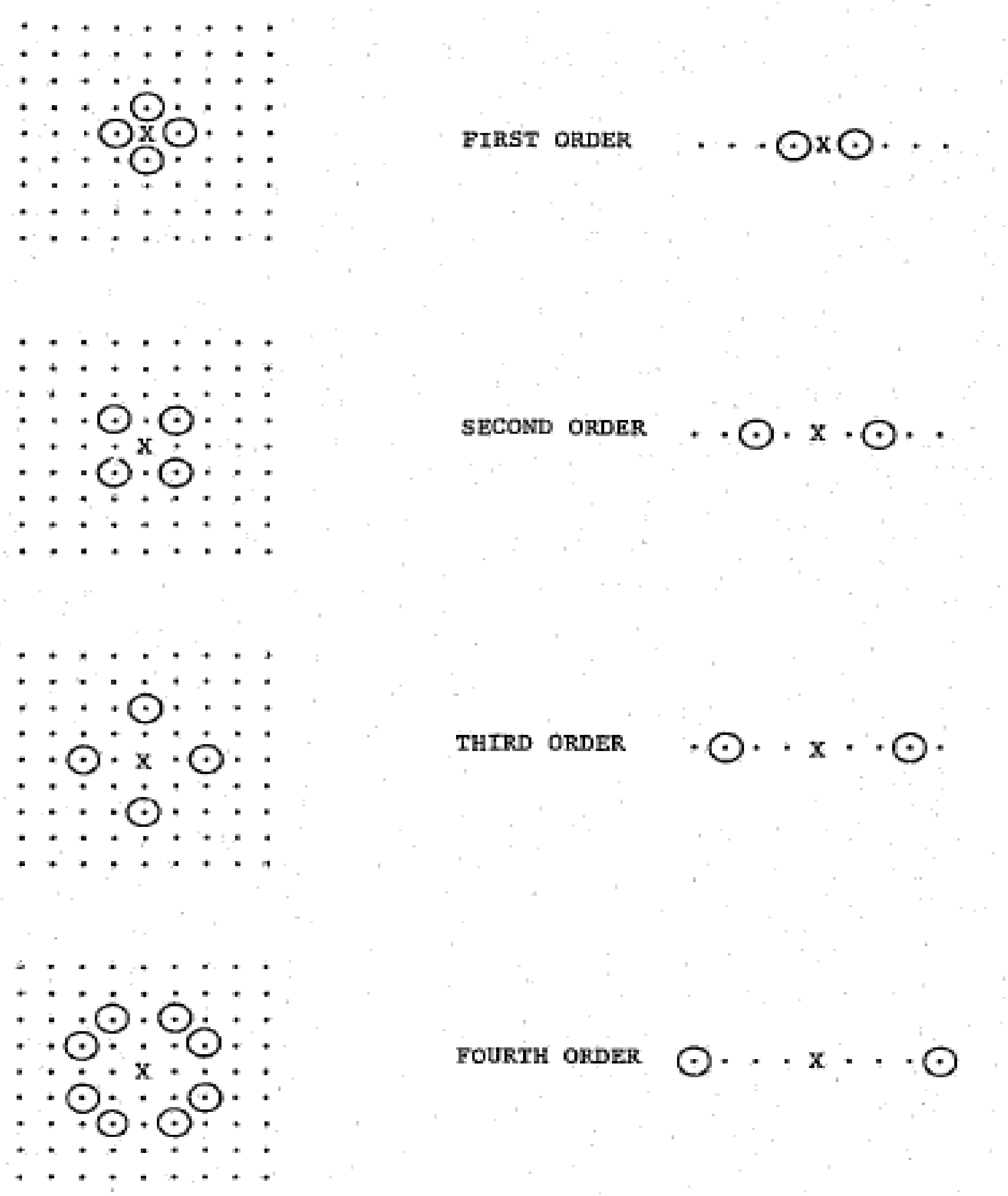


Fig.1: Schematic representation of spatial weight grid

Be that as it may, in spatial weight matrix, row normalization is a common practice i.e. making all rows sum to one is common practice. However, in some studies, column normalization has been used, allowing the matrix to represent influence exerted by *i* rather than accepted influence from *j*. The choice of weighting scheme is nontrivial and can be very important because different weight matrices often lead to different inferences being drawn and can introduce bias into an analysis. In spatiotemporal data, if the relative contributions of the spatial neighbors of a unit remain the same across all times may not be reasonable. These weights, in any case, must reflect a hierarchical ordering of spatial neighbors. First order neighbors are those which are closest to the chosen site. Second order neighbors are farther away than first order neighbor's, yet closer than third order neighbors. The schematic representation of spatial weight grid is represented (Pfeifer and Deutsch, 1980b) in figure 1.

1.1.4. STARMA Modeling Procedure

As like Box-Jenkins univariate ARIMA methodology the STARMA model is also build by three stage procedure of model building viz., identification, estimation and diagnostic checking, proposed by Pfeifer and Deutsch (1980b). The STARMA model is said to stationary if covariance structure of Z(t) does not change with time and every Z(t) lie inside the unit root circle i.e. the STAR model are invertible and STAMA models are stationary.

1.1.4.1.Model Identification

The space time autocorrelation function (STACF) and space time partial autocorrelation function (STPACF) are used to identify the STAR and STMA order. Like univariate ARIMA model, the STAR and STMA model orders are identified in view of significant STAR and STMA spikes. The space time autocorrelation function (STACF) between lth and k^{th} order neighbor's s time lag apart (s=1,…,k and h=0,1,…,λ) is given underneath;

$$\rho_{lk}(s) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-S} W^{(l)} Z_i(t) W^{(k)} Z_i(t+s)}{\left[\sum_{i=1}^{N}\sum_{t=1}^{T-S} (W^{(l)} Z_i(t))^2 \ . \ (W^{(k)} Z_i(t+s))^2\right]^{\frac{1}{2}}} \qquad \ldots(1.4)$$

The space time partial autocorrelation function (STPACF) is expressed in following equation;

$$\rho_{h0}(s) = \sum_{j=1}^{k}\sum_{l=0}^{\lambda} \phi_{jl}\rho_{hl}(s-j) \qquad \ldots(1.5)$$

Characteristics of the theoretical space-time autocorrelation and partial autocorrelation functions for STAR, STMA and STARMA models (1.1) are depicted in following table.

Table 1: STACF and STPACF of STAR, STMA and STARMA models

Process	**STACF**	**STPACF**
STAR	tails off with both space and time	cuts off after p lags in time and λ_p lags in space
STMA	cuts off after q lags in time and m_q lags in space	tails off with both space and time
STARMA	tails off	tails off

1.1.4.2.Model Parameter Estimation

The maximum likelihood estimates of

$\boldsymbol{\Phi} = [\phi_{10,}\phi_{11}, \dots, \phi_{1\lambda_1}, \dots, \phi_{p0}, \phi_{p1}, \dots, \phi_{p\lambda_p}]'$ and

$\Theta = [\theta_{10,}\theta_{11}, \dots, \theta_{1\lambda_1}, \dots, \theta_{q0}, \theta_{q1}, \dots, \theta_{p\lambda_q}]'$ rely on the assumption of errors i.e. which are normally distributed with mean zero and variance-covariance matrix equal to $\sigma^2 I_N$. The likelihood function for the same is defined as follows;

$$f(\varepsilon|\Phi,\Theta,\sigma^2)=(2\pi)^{\frac{-TN}{2}}|\sigma^2 I_{NT}|^{\frac{-1}{2}}\exp\{-\frac{1}{2\sigma^2}\epsilon' I\epsilon\}$$

$$=(2\pi)^{\frac{-TN}{2}}(\sigma^2)^{\frac{-TN}{2}}\exp\{-\frac{S(\Phi,\Theta)}{2\sigma^2}\} \qquad \dots(1.6)$$

Where,

$S(\Phi, \Theta) = \epsilon' I\epsilon = \sum_{i=1}^{N}\sum_{t=0}^{T}\epsilon_i^2(t)$ is the sum of squares of the errors and $\epsilon' = [\epsilon_1(1),\dots, \epsilon_1(T), \dots, \epsilon_N(1), \dots, \epsilon_N(T)]$. Finding the values of the parameters that maximize the likelihood function is equivalent to finding the values of Φ and Θ that minimize the sum of squares $S(\Phi, \Theta)$. Therefore, the problem is reduced to finding the least squares estimates of Φ and Θ .

The errors $\varepsilon(t)$ need to be recursively calculated using the equation:

$$\varepsilon(t)=z(t) + \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}\,W^{(l)}z(t-k) - \sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}\,W^{(l)}\varepsilon(t-k) \qquad \dots(1.7)$$

for t = 1, ..., T and for given values of the parameters (Φ,Θ).

Because the values of the observations z and of the errors care unknown for times before time 1, these initial values need to be calculated. Thus, for any given choice of the parameters (Φ,Θ) and starting values $(z *, c *)$ the set of values $c(cI >, e\ I\ z *, c *, W)$ could be calculated successively given a data set z. The log likelihood associated with the parameter values (Φ, Θ, σ^2) conditional on the choice of $(z *, c *)$ would be

$$l_*(\Phi, \Theta, \sigma^2) = -\frac{TN}{2}\ln(2\pi) - \frac{TN}{2}\sigma^2 - \frac{S_*(\Phi,\Theta)}{2\sigma^2} \qquad \ldots(1.8)$$

So, for fixed σ^2 , the conditional maximum likelihood estimates of Φ, Θ are the conditional least squares estimates obtained by finding the values of Φ, Θ that minimize the conditional sum of squares function

$$S_*(\Phi, \Theta) = \sum_{i=1}^{N}\sum_{t=0}^{T}\epsilon_i^2(t) \qquad \ldots(1.9)$$

1.1.4.3.Diagnostic-Checking

At this stage the objective is to determine if the model does adequately represent the data. If the fitted model adequately represents the data, the residuals should be gaussian white noise, i.e., should be distributed normally with mean zero and variance-covariance matrix equal to $\sigma^2 I_N$. One way of testing for correlation is to calculate the sample space-time auto correlations of the residuals and check for additional significant structure. If the model is adequate then,

$$var\big(\hat{\rho}_{l0}(s)\big) \approx \frac{1}{N(T-s)} \qquad \ldots(1.10)$$

Where $\hat{\rho}_{l0}(s)$ is the space-time autocorrelation function of the residuals of the fitted model. Thus, the residual space-time autocorrelations, since they are approximately normal, can be standardized and checked for significance. If the residuals are not independent the pattern is identified, and the tentative model updated.

Case Study: Modeling and Forecasting of monthly mean maximum temperature of nine districts of north Karnataka. **Rathod et al (2018).**

In this study monthly mean maximum temperature of nine districts of north Karnataka state of India (Fig. 1) are considered to model and forecast using the proposed STARMA methodologies. The data from January 2000 to August, 2015 has been utilized for model building and data from September, 2015 to August, 2016 used for model validation (Forecasting performance).

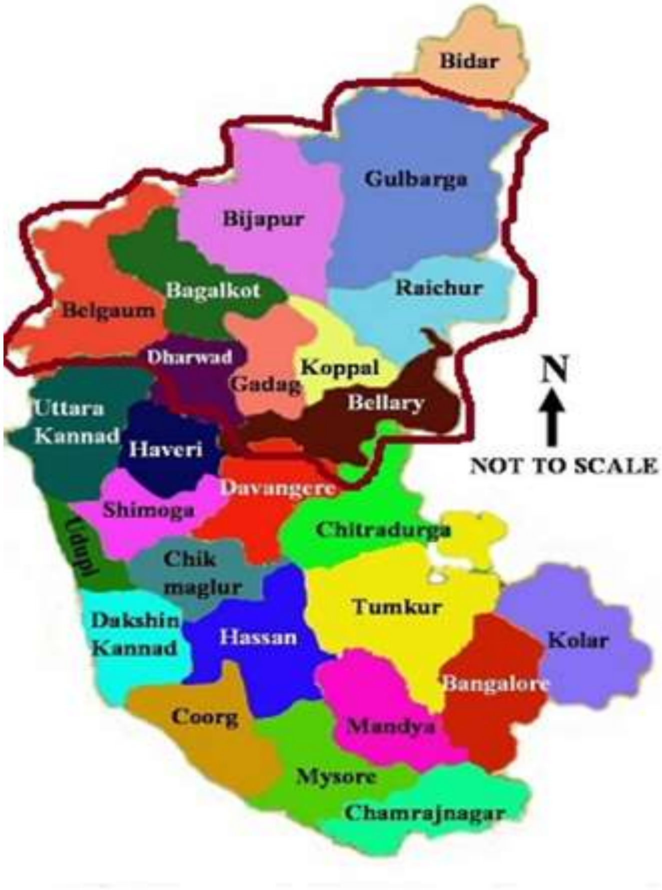
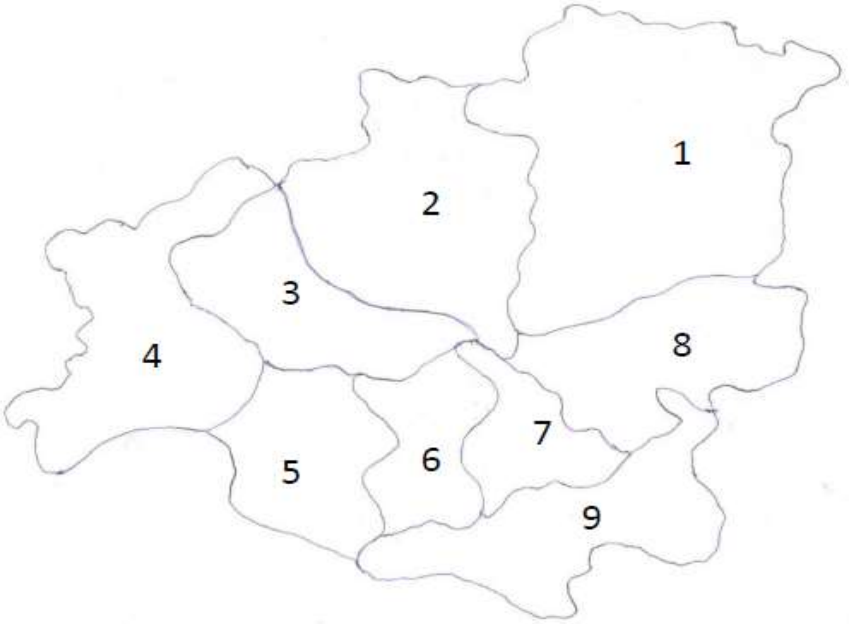


Fig. 1. Geographical map of karnataka

Construction of spatial weight matrix:

As explained in methodology section, the spatial weight matrix has been constructed by assigning equal weightage to each neighbor. The map of nine locations under consideration is delineated in figure 2.10 and each location are represented by numbers from one to nine. Considering the neighboring locations, connectivity spatial weight matrices have been considered. For instance, for location 1, location 2 and location 8 are first order neighbors. Again, 3, 6 and 7 are second order neighbors to location one. In a similar manner, first and second order neighbors for all nine locations are reported in table 1. Considering the numbers of neighbors, the spatial weights have been doled out to each location. In uniform spatial weight matrix equal weights are relegate to each neighbors. To make row normalization i.e. making all rows sum to one we divide, one by number of neighbors i.e. $\frac{1}{n}$, here n is number of neighbors. For example, for first location (Gulbarga) there are two first order neighbors, then we divide one by two and assign 0.5 as weight to each locations. As we calculated weight for first location, one can proceed in same manner to calculate weights for all nine locations. In light of this procedure first order spatial weight matrix has been calculated in table 3. In this work attempt has been made to incorporate second order spatial weight matrix in STARMA model. For first location 3, 6 and 7 are second order neighbors, then we divide one by three and assign 0.33 as weight to each location; in the same manner one can proceed further to calculate weights to all nine locations for second order neighbors. The second order spatial weight matrix for all nine locations are depicted in table 4. To compute STACF and STPACF, zero order (Table 2) first order and second order spatial weight matrix (Table 3 and 4) is needing to be incorporate in the model. In first order spatial weight matrix, since we do not assign weights to any neighbors, diagonal elements end up noticeably equal to one.

Fig. 2.: Map of districts/locations considered



1. Gulbarga
2. Bijapur
3. Bagalkot
4. Belgaum
5. Dharwad
6. Gadag
7. Koppal
8. Raichur
9. Bellary

Table 1: Neighbors of each site for each spatial order

Location	Order	
	1	2
1	2,8	3,6,7
2	1,3	4,8,7
3	2,4,5,6	7,8
4	3,5	2
5	3,4,6	9
6	3,5,7,9	2,8
7	6,8,9	1,2,3
8	1,7,9	2,3,6
9	6,7,8	5

Table 2.: Spatial weight matrix of order zero

Location	Gulbarga	Bijapur	Raichur	Bagalkot	Belgaum	Dharwad	Gadag	Koppal	Bellary
Gulbarga	1	0	0	0	0	0	0	0	0
Bijapur	0	1	0	0	0	0	0	0	0
Raichur	0	0	1	0	0	0	0	0	0
Bagalkot	0	0	0	1	0	0	0	0	0
Belgaum	0	0	0	0	1	0	0	0	0

Dharwad	0	0	0	0	0	1	0	0	0
Gadag	0	0	0	0	0	0	1	0	0
Koppal	0	0	0	0	0	0	0	1	0
Bellary	0	0	0	0	0	0	0	0	1

Table 3: First order spatial weight matrix for Maximum temperature data

Location	Gulbarga	Bijapur	Raichur	Bagalkot	Belgaum	Dharwad	Gadag	Koppal	Bellary
Gulbarga	0	0.5	0	0	0	0	0	0.5	0
Bijapur	0.5	0	0.5	0	0	0	0	0	0
Raichur	0	0.25	0	0.25	0.25	0.25	0	0	0
Bagalkot	0	0	0.5	0	0.5	0	0	0	0
Belgaum	0	0	0.33	0.33	0	0.33	0	0	0
Dharwad	0	0	0.25	0	0.25	0	0.25	0	0.25
Gadag	0	0	0	0	0	0.33	0	0.33	0.33
Koppal	0.33	0	0	0	0	0	0.33	0	0.33
Bellary	0	0	0	0	0	0.33	0.33	0.33	0

Table 4: Second order spatial weight matrix

Location	Gulbarga	Bijapur	Raichur	Bagalkot	Belgaum	Dharwad	Gadag	Koppal	Bellary
Gulbarga	0	0	0.33	0	0	0.33	0.33	0	0
Bijapur	0	0	0	0.33	0	0	0.33	0.33	0
Raichur	0	0	0	0	0	0	0.5	0.5	0
Bagalkot	0	1	0	0	0	0	0	0	0
Belgaum	0	0	0	0	0	0	0	0	1
Dharwad	0	0.5	0	0	0	0	0	0.5	0
Gadag	0.33	0.33	0.33	0	0	0	0	0	0
Koppal	0	0.33	0.33	0	0	0.33	0	0	0
Bellary	0	0	0	0	1	0	0	0	0

STARMA model fitting:

In this article STARMA model was estimated using the three-stage procedure explained by Pfeiffer and Deutsch (Pfeiffer and Deutsch, 1980a). As explained in methodology section, STARMA estimation procedure is extension of Box-Jenkins ARIMA methodology in spatio-temporal set up. As like ARMA It likewise has three stages of model building *viz.*, model identification, estimation and diagnostic checking. Considering the significant spikes in STACF and STPACF plots, the model order STARMA (1 0 1), has been identified. Parameters of the identified models are estimated using maximum likelihood method and are given in table 5, alongside their standard errors and probability values. The estimated parameters are then consolidated in the model and predicted values were acquired. For diagnostic checking Multivariate Box-Pierce Non-Correlation test has been applied and the residuals are observed to be non-correlated. Further, performance of models under consideration is depicted in table 6.

Table 5: STARMA Model parameters

Spatial lag	Slag 0		Slag 1		Slag 2	
	AR	MA	AR	MA	AR	MA
Parameters	-0.66 (0.023)	0.119 (0.010)	0.171 (0.052)	0.213 (0.0157)	0.79 (0.089)	0.11 (0.116)
Probability	<0.001	<0.001	0.013	0.004	<0.001	0.010

Multivariate Box-Pierce Non-Correlation Test of residuals: Chi-square=69.86 (p=0.31)
Values in the parenthesis indicates the standard error

The mean absolute percentage error (MAPE) has been computed to compare the forecasting performance of ARIMA and STARMA model (Table 6). In view of the lowest MAPE value of proposed STARMA model for all the locations, it is affirmed that STARMA model outflanked the Box-Jenkins ARIMA model in all the locations.

Table 6: Modeling Performance in terms of MAPE

Sl. No	Location	ARIMA	STARMA
1	Gulbarga	2.54	1.30
2	Bijapur	2.73	1.29
3	Raichur	2.36	1.24
4	Bagalkot	2.80	1.49
5	Belgaum	3.42	2.07
6	Dharwad	3.31	1.69
7	Gadag	2.97	1.56
8	Koppal	2.89	1.41
9	Bellary	2.45	1.24

R scripts to implement STARMA and ARIMA model

```
install.packages(starma)
install.packages(spdep)
install.packages(lmtest)
install.packages(forecast)
install.packages(fNonlinear)
library(starma)
library(spdep)
library(lmtest)
library(forecast)
library(fNonlinear)
w0.mat<-as.matrix(read.table(file.choose(),header=TRUE))
w1.mat<-as.matrix(read.table(file.choose(),header=TRUE))
W <- list(order0=w0.mat, order1=w1.mat)
st<-as.matrix(read.table(file.choose(),header=TRUE))  # data read
st<- stcenter(st)
stacf(st, W, tlag.max=36)
stpacf(st, W, tlag.max=36)
# model fitting
starma(st, wlist = W, ar = 1, ma = 1)
summary(st.fit)
st1=read.table(file.choose(), header = T)  # import again
st11=st1$EngNE
st11.fit=auto.arima(st11)     # arima fitting
accuracy(st11.fit)
st11.fit
coeftest(st11.fit)
res.st=st.fit$residuals  # residuls of STARMA model
resdata=data.frame(res.st)  # create data frame
write.csv(as.data.frame(res.st), file="stres.csv")
st12=st1$EA              # second location arima fitting
st11f=auto.arima(st12)
st11f=auto.arima(st11)
accuracy(st11.fit)
```

Suggested Readings:

Box, G.E.P. and Jenkins, G. (1970). Time series analysis, Forecasting and control, Holden-Day, San Francisco, CA.

Ding, Q., X. Wang, X. Zhang, and Z. Sun. (2011). Forecasting Traffic Volume with Space–Time ARIMA Model. Advanced Materials Research, 156–57, 979–83.

Pfeifer, P.E., and Bodily, S.E. (1990). A test of space-time ARMA modeling and forecasting with an application to real estate prices, International Journal of Forecasting, 16, 255-272.

Pfeifer, P.E., and Deutsch, S.J. (1980). A Comparison of Estimation Procedures for the Parameters of the STAR Model. Communication in Statistics, simulation and Comput., B9(3), 255-270.

Pfeifer, P.E., and Deutsch, S.J. (1980a). A three-stage iterative procedure for space-time modeling. Technometrics, 22(1), 35-47.

Pfeifer, P.E., and Deutsch, S.J. (1981). Variance of the Sample-Time Autocorrelation Function of Contemporaneously Correlated Variables. SIAM Journal of Applied Mathematics, Series A, 40(1), 133-136.

Rathod, S., Gurung,B., Singh, K.N. and Ray, M. (2018). An improved Space- time Autoregressive Moving Average (STARMA) model for Modelling and Forecasting of Spatio-Temporal time-series data. Journal of the Indian Society of Agricultural Statistics. 72(3): 239-253.

Nonlinear Growth Model: Introduction and overview

[1]Mrinmoy Ray, [1]K N Singh, [1]Achal Lama, [1]Kanchan Sinha and [2]Santosha Rathod
[1]ICAR-IASRI, New Delhi
[2]ICAR-IIRR, Hyderabad

1. Introduction

Growth is defined as an “Irreversible increase in size and volume and is the consequence of differentiation and distribution occurring in the plant/animal”. A model is a schematic representation of the conception of a system or an act of mimicry or a set of equations, which represents the behaviour of a system. Also, a model is “A representation of an object, system or idea in some for other than that of the entity itself”. Its purpose is usually to aid in explaining, understanding or improving performance of a system.

TYPES OF MODELS

Depending upon the purpose for which it is designed the models are classified into different groups or types. Of them a few are:

a. Statistical models: These models express the relationship between yield or yield components and weather parameters. In these models relationships are measured in a system using statistical techniques (Table 1).

Example: Step down regressions, correlation, etc.

b. Mechanistic models: These models explain not only the relationship between weather parameters and yield, but also the mechanism of these models (explains the relationship of influencing dependent variables). These models are based on physical selection.

c. Deterministic models: These models estimate the exact value of the yield or dependent variable. These models also have defined coefficients.

d. Stochastic models: A probability element is attached to each output. Foreach set of inputs different outputs are given along with probabilities. These models define yield or state of dependent variable at a given rate.

e. Dynamic models: Time is included as a variable. Both dependent and independent variables are having values which remain constant over a given period of time.

f. Static: Time is not included as a variable. Dependent and independent variables having values remain constant over a given period of time.

g. Simulation models: Computer models, in general, are a mathematical representation of a real world system. One of the main goals of crop simulation models is to estimate agricultural production as a function of weather and soil conditions as well as crop management. These models use one or more sets of differential equations, and calculate both rate and state variables over time, normally from planting until harvest maturity or final harvest.

Statistical Modelling

A fundamental problem in statistics is to develop models based on a sample of observations and inferences using the model so developed. In almost all branches of agriculture including animal sciences and fisheries, vast amounts of data pertaining to production/productivity of various crops, and import-export of various agricultural commodities, etc. are being collected sequentially over time. One characteristic of such data is that the successive observations are dependent. Each observation of the observed data series, Y_t, may be considered as a realization of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t, t \in T\}$, where $T = \{ 0, \pm 1, \pm 2, \ldots\}$, and apply standard time-series approach to develop an ideal model which will adequately represent the set of realizations and also their statistical relationships in a satisfactory manner. Forecasting of these types of time-series data is of great importance for planners and policy makers. During the last some decades, a new area of “Nonlinear time-series modelling” has rapidly been developing. Here, there are basically two possibilities, viz. Parametric or Nonparametric approaches. Evidently, if in a particular situation, we are quite sure about the functional form, we should use the former, otherwise the latter may be employed.

Parametric and Nonparametric Approaches

Over the last several decades, regression analysis has become increasingly popular as a tool for statistical modelling and data analysis. This provides information on relationship between a response variable and one or more predictor variables. The main objective is to express the mean of response as a function of predictor variables. General regression model is of the form

$$Y \quad = \quad m\,(X) + \;\varepsilon$$

where$Y$ is the response variable, $m(X) = E\,(Y|\,X)$ is the mean response or regression function and ε is the error. The regression function $m(X)$ is usually unknown and the objective is to obtain a suitable estimator of $m(X)$ using a sample of observations.

In the linear regression, it is assumed that the mean of the response variable Y is a linear function of predictor variable(s) X of the form

$$E\,(Y|\,X) \;\; = \; X\beta$$

i.e. $m(X)$ is linear in parameters. The parameter vector β is usually estimated by the Method of least squares. In nonlinear regression, it is assumed that the mean of the response variable is a nonlinear function of the predictor variable (s) X of the form

$E(Y|X)=m(X,\beta)$

i.e. $m(X)$ is nonlinear in parameters. Generally, there will be no closed form expression for the estimates of β and iterative procedures are required for estimation of parameters.

A parametric regression model (linear or nonlinear) assumes that the form of m is known except for some unknown parameters, and shape of the regression function is entirely dependent on the parameters. Often, it is difficult to guess the most appropriate functional form just from looking at the data. Sometimes there may not be some suitable parametric form to express the regression function. In such situations, the nonparametric regression approach, which does not require strong assumptions about the shape of the regression function, is very useful. A nonparametric regression model only assumes that *m* belongs to some infinite dimensional collection of functions. One limitation of above approach is that it generally relies upon certain assumptions about smoothness of the function being estimated, which may not hold in reality. This may result in over smoothing of the data under consideration.

LINEAR MODEL

A mathematical model is an equation or a set of equations which represents the behaviour of a system. It can be either 'linear' or 'nonlinear'. A linear model is one in which all the parameters appear linearly.

NONLINEAR MODELS

Any type of statistical inquiry in which principles from some body of knowledge enter seriously into the analysis is likely to lead to a 'Nonlinear model'. Such models play a very important role in understanding the complex inter-relationships among variables. A 'nonlinear model' is one in which at least one of the parameters appears nonlinearly. More formally, in a 'nonlinear model', at least one derivative with respect to a parameter should involve that parameter.

- Examples of a nonlinear model are:

$$Y(t) = \exp(at+bt^2) \qquad (1a)$$

$$Y(t) = at + \exp(-bt) \qquad (1b)$$

Note. Some authors use the term 'intrinsically nonlinear' to indicate a nonlinear model which can be transformed to a linear model by means of some transformation.

For example, the model given by Eq. (1a) is 'intrinsically nonlinear' in view of the transformation $X(t) = \log_e Y(t)$.

a. MALTHUS MODEL:

In 1798 the Englishman Thomas R. Malthus posited a mathematical model of population growth. His model, though simple, has become a basis for most future modeling of biological populations. His essay, "An Essay on the Principle of Population," contains an excellent discussion of the caveats of mathematical modeling and should be required reading for all serious students of the discipline. Malthus's observation was that, unchecked by environmental or social constraints, it appeared that human populations doubled every twenty-five years, regardless of the initial population size. Said another way, he posited that populations increased by a fixed proportion over a given period of time and that, absent constraints, this proportion was not affected by the size of the population. By way of example, according to Malthus, if a population of 100 individuals increased to a population 135 individuals over the course of, say, five years, then a population of 1000 individuals would increase to 1350 individuals over the same period of time. Malthus's model is an example of a model with one *variable* and one *parameter.* A variable is the quantity we are interested in observing. They usually change over time. Parameters are quantities which are known to the modeller before the model is constructed. Often they are constants, although it *is* possible for a parameter to change over time. In the Malthusian model the variable is the population and the parameter is the population growth rate.

If N(t) denotes the population size or biomass at time t and r is the intrinsic growth rate, then the rate of growth of population size is given by
$dN/dt = rN$

Therefore, $N(t) = N_o\exp(rt)$

Note : Malthus model can be used for describing growth of simplistic organisms, which begin to grow by binary splitting of cells.

Drawback:$N(t)\to\infty$ as $t\to\infty$, which cannot happen in reality.

Malthus hypothesized that unchecked population growth would quickly exceed carrying capacity, leading to overpopulation and social problems**.**

Note. The parameter ris assumed to be positive.

b. MONOMOLECULAR MODEL:

The monomolecular model assumes a carrying capacity of one, that is, the maximum level of disease is one, so disease severity or incidence is measured as a proportion. Diseased plant tissue may only lie between zero (healthy) and one (complete disease). It also assumes the absolute rate of change is proportional to the healthy tissue i.e., (1-*y*). After creating the plot mono function and trying the example set of parameter values, try replacing the parameter values with others to see how the shape of the relationship changes.

It describes progress of growth in which it is believed that the rate of growth at any time is proportional to the resources yet to be achieved i.e.

$dN/dt = r(K-N)$,

where K is the carrying capacity.

or $N(t) = K- (K-N_o) \exp(-rt)$

Drawback: No point of inflexion.

c. LOGISTIC MODEL:

Logistic model was developed by Belgian mathematician Pierre Verhulst (1838) who suggested that the rate of population increase may be limited, i.e., it may depend on population density. Population growth rate declines with population numbers, N, and reaches 0 when N = K. Parameter K is the upper limit of population growth and it is called carrying capacity. It is usually interpreted as the amount of resources expressed in the number of organisms that can be supported by these resources. If population numbers exceed K, then population growth rate becomes negative and population numbers decline. This model is represented by the differential equation:

$$dN/dt = rN (1-N/K) \qquad (1)$$

Therefore, $N(t) = K/[1+(K/N_o-1) \exp(-rt)]$. The graph of N(t) versus t is elongated S-shaped and the curve is symmetrical about its point of inflexion.

d. GOMPERTZ MODEL

This is another model having a sigmoid type of behaviour and is found to be quite useful in biological work. The Gompertz curve was originally derived to estimate human mortality by Benjamin Gompertz (Gompertz, B. "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life

Contingencies." Phil. Trans. Roy. Soc. London 123, 513-585, 1832). Charles Winsor (1932) presents an early description of the use of this equation to describe growth processes. However, unlike the logistic model, this is not symmetric about its point of inflexion.

The differential equation for this model is

$dN/dt = rN \log_e (K/N)$ (2)

or $N(t) = K \exp[\log_e (N_o / K) \exp(-rt)]$

e. **RICHARDS MODEL:**

The Richards curve or generalized logistic is a widely used growth model that will fit a wide range of S-shaped growth curves. There are both 4 and 5 parameter versions in common use. The logistic curve is symmetrical about the point of inflection of the curve. To deal with situations where the growth curve is asymmetrical, Richards (1959) added an additional parameter

This model is given by

$$N(t) = K N_o / [N_o + (K^m - N_o^m) \exp(-rt)]^{1/m} \quad . \qquad (4)$$

However, unlike the earlier models, this model has four parameters.

Drawback. Number of parameters is more.

f. **MIXED-INFLUENCE MODEL:**

This is a mixture of ‘Monomolecular’ and ‘Logistic’ Models. It is given by

$dN/dt = r (K-N) +s N (1-N/K)$,

FITTING OF NONLINEAR MODELS

The above models have been posed deterministically. Obviously this is unrealistic and so we replace these deterministic models by statistical models by adding an error term on the right hand side and making appropriate assumptions about them. This results in a ‘Nonlinear statistical model’. As in linear regression, in non-linear case also, parameter estimates can be obtained by the ‘Method of least squares’. However, minimization of residual sum of squares yield normal equations which are nonlinear in the parameters. Since it is not possible to solve nonlinear equations exactly, the next alternative is to obtain approximate analytic solutions by employing iterative procedures.

- Four main methods of this kind are:
  i) Linearization (or Taylor Series) method

  ii) Steepest Descent method

  iii) Levenberg-Marquardt's method

  iv) Do not use Derivatives method

The details of these methods along with their merits and demerits are given in Draper and Smith (1998). Neither the Linearization method nor the Steepest descent method is ideal. The most widely used method of computing nonlinear least squares estimates is the Levenberg-Marquardt's method. This method represents a compromise between the other two methods and combines successfully the best features of both and avoids their serious disadvantages. It is good in the sense that it almost always converges and does not 'slow down' at the latter part of the iterative process.

- SPSS package has NLR option, while SAS package has NLIN option to fit nonlinear statistical models based on Levenberg-Marquardt algorithm.
- Most important thing is the 'Meaningful interpretation' of parameter estimates.

CHOICE OF INITIAL VALUES

All the procedures for nonlinear estimation require initial values of the parameters and the choice of good initial values is very crucial. However, there is no standard procedure for getting initial estimates. The most obvious method for making initial guesses is by the use of prior information. Estimates calculated from previous experiments, known values for similar systems, values computed from theoretical considerations all these form ideal initial guesses.

Some other methods are:

(i) Linearization:

After ignoring the error term, check the form of the model to see if it could be transformed into a linear form by means of some transformation. In such cases, linear regression can be used to obtain initial values.

(ii) Solving a system of equations:

If there are p parameters, substitute for p sets of observations into the model ignoring the error. Solve these equations for the parameters, if possible. Widely separated x_i often work best.

R code

Monomolecular growth model

```
 z=read.csv(file.choose(), header=TRUE)
head(z)
kk=data.frame(z)
grz1=nls(y~k-(k-y0)*exp(-r*t),data=kk,  start=list(k=1 ,y0=0.03,r=0.1))
summary(grz1)
 fitted=kk$y-resid(grz1)
kkk=data.frame(fitted)
MSE.nn <- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors <- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,1),axes=FALSE, ann=FALSE)
axis(1, at=1:20, lab=c(0:19))
axis(2, las=1, at=0.2*0:5)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22, lty=1:2);
zz=resid(grz1)
predicted= 0.99651-(0.99651-0.08844)*exp(-0.26727*20)
```

Gompertz model

```
z=read.csv(file.choose(), header=TRUE)
 head(z)
 kk=data.frame(z)
gr1=nls(y~k*exp(log(y0/k)* exp(-r*t)),data=kk,  start=list(k=50,y0=11.72,r=0.1))
summary(gr1)
fitted=kk$y-resid(gr1)
kkk=data.frame(fitted)
MSE.nn <- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors <- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,35),axes=FALSE, ann=FALSE)
axis(1, at=1:38, lab=c(0:37))
axis(2, las=1, at=5*0:8)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22, lty=1:2);
```

logistic model

```
z=read.csv(file.choose(), header=TRUE)
 head(z)
 kk=data.frame(z)
gr2=nls(y~k/(1+(k/y0-1)* exp(-r*t)), data=kk,  start=list(k=50,y0=11.72,r=0.1))
summary(gr2)
fitted=kk$y-resid(gr2)
kkk=data.frame(fitted)
MSE.nn <- sum((kk$y- kkk)^2)/nrow(kkk)
plot_colors <- c("blue","red")
plot(kk$y,type="o", col=plot_colors[1], ylim=c(0,35),axes=FALSE, ann=FALSE)
axis(1, at=1:38, lab=c(0:37))
axis(2, las=1, at=5*0:8)
box()
lines(fitted,type="o", pch=22, lty=2,col=plot_colors[2])
title(main="Actual vs predicted",col.main="red", font.main=4)
title(xlab= "Time", col.lab=rgb(0,0.5,0))
title(ylab= "Growth", col.lab=rgb(0,0.5,0))
legend("topleft",c("actual", "predicted"),cex=0.8, col=plot_colors, pch=21:22, lty=1:2);
```

An Introduction to Fuzzy Set and Fuzzy Time Series Forecasting

Amit Saha[1], K. N. Singh[2], Mrinmoy Ray[2] and Sanjay Kumar[3]
[1]Central Silk Board, Ministry of Textiles, Government of India
[2]ICAR-IASRI, New Delhi
[3]Govind Ballabh Pant University of Agriculture and Technology, Pantnagar

amits.csb@gov.in

1. Introduction:

Time series forecasting is well known method of forecasting in many areas. Time series forecasting is popular because of easiness of evaluation of time series data for getting the forecast values. Another reason of its' popularity is that the real world data are mostly time series data. In time series forecasting, time series data are taken as a crisp value. However, data may not be precise and complete in all the cases, e.g. water level data of river, temperature data etc. Fuzzy techniques are appropriate in those cases when vagueness has been seen in the data. Fuzzy data can be found in artificial intelligence, quality control, biology, psychometry, agriculture, social economy, image recognition etc. Fuzzy time series model can improve the utilization of the data.

2. Situations where fuzzy techniques are useful?

Fuzzy techniques are applied in those conditions where-

1. People's decisions are involved.
2. When the data are imprecise.
3. Assumptions of distribution are not satisfied.
4. The past data are less in which we cannot use the existing time series model.

3. Fuzzy Set:

3.1 Fuzzy set theory:

Zadeh in 1965 defined fuzzy set as "A fuzzy set is a class of objects with a continuum of grades of membership". He discussed the various properties of fuzzy sets. Various operations on fuzzy set such as union, intersection and complement were discussed which are different from the basic operations of conventional set. In conventional set, an element can be either the member or nonmember of a specific set. Our answer will be always in Yes or No in respect to presence or non-presence of a particular element in a set. Let, we take a set of overweight persons, all of them are either overweight or not. Nothing is here in the middle of them. This is the main drawback of the conventional set. In real world, it is not always possible to describe some concepts by their presence or non-presence in a particular set. Suppose, a person can be recognized as overweight person if his weight is 100 kg or more and he is not overweight if his weight is less than 70kg. But, it does not give the answer whose weight lies in between them. So, there is some haziness where someone cannot clearly say either yes or no. But, fuzzy set uses some membership function to assign various degrees of membership to the elements. So that, the shifting from yes to no is continuing process rather than sudden jump. Generally, the degree of membership lies between 0 and 1. The belongings of an element to a specific set is high when degree of membership is more.

Definition: A fuzzy set is defined as the set of ordered pairs in which every single pair (fuzzy singleton) contains an element and the value of its belonginess to the fuzzy set. Mathematically, fuzzy set can be written as:-

$$F = \left[\left(x_{i,}\mu_F\left(x_{i,}\right)\right)\right] \qquad (1)$$

Where, $i = 1,2,3 \ldots n$, x_i is the i^{th} element which is a member of F and $\mu_F\left(x_{i,}\right)$ represents the degree of membership of x_i in F. Fuzzy set can also be written as:

$F = I_1/x_1 + I_2/x_2 + \ldots I_n/x_n$ (Only for discrete and finite fuzzy set)

Where, I_1, I_2,…, I_n are the intervals of the defined fuzzy set and the above operations are not algebric operations.

3.2 *Linguistic variable*: Linguistic variables are quantitative variable which are used to represent the fuzzy numbers. The values of linguistic variable are sentences or words rather than numeric.

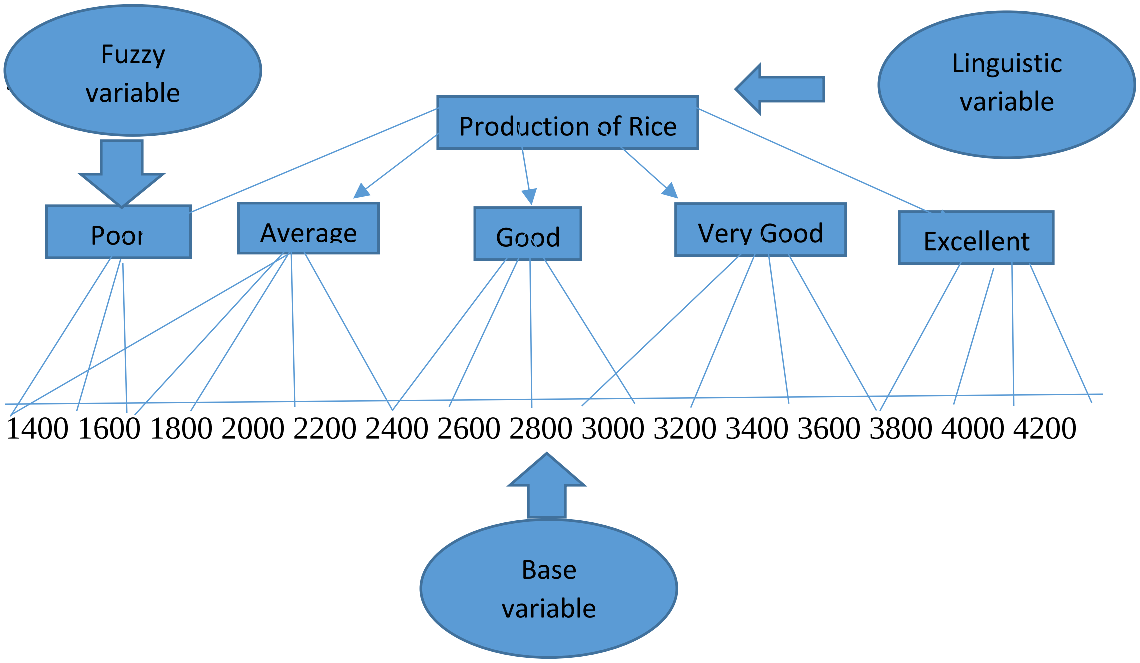


Figure 1: Illustration of linguistic variables and fuzzy

For example, if we take the height of a person and represent it with linguistics expressions as very much tall, very tall, tall, more than medium tall, medium tall, short, very short and very much short. These linguistic variables are associated with any one of various types of membership function. The range of the different linguistic variables will be defined by the selected membership function. Fig. 2 depicts a pictorial representation of fuzzy and linguistic variable.

3.3 *Membership Function:*

Zadeh first introduced the membership functions in his research paper “Fuzzy Sets” which was published in Information and Control journal in the year of 1965. The fuzziness of the data is described by the membership function. It means that membership function provides the degree of membership to the various element in the fuzzy set. There are many membership functions available in literature. The use of a particular type of membership function depends on the concept as well as the context to be demonstrated. Membership functions can be depicted through the graphical representation. The different graphs of the membership functions have different properties along with the diverse shapes. The various membership functions are:

1. Triangular membership functions.
2. Trapezoidal membership functions.
3. S membership functions.

1. Triangular membership functions: It is defined by three parameters and this function is a piecewise linear function. Saha *et al.* (2019) used the triangular membership functions in space-time series modelling and forecasting. Mathematically, it can be written as:-

$$w_F = \begin{cases} \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & x > c \end{cases} \tag{2}$$

2. Trapezoidal membership Function: It is defined by four parameters. Mathematically, this function is written as:-

$$w_F = \begin{cases} \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c}, & c \leq x \leq d \\ 0, & x > d \end{cases} \tag{3}$$

3. S membership function: It is defined by two parameters and can be written as:

$$w_F(x; a, b) = \frac{1}{1+e^{-b(x-a)}} \tag{4}$$

Where, the midpoint and slope value is denoted by a and b respectively and b must have the positive value. Another important thing of sigmoidal membership function is that it never goes to 0 or 1.

Other membership functions are G And L open shoulder membership functions, Bell shaped Function, Z function etc. A graphical representation has been provided below to explain the membership function.

Example (Fig. 2): Suppose, a person is tall if his height is 180 cm or more. That means, If the height of Ratan is ≥180 cm , then he is a tall person, otherwise not. Suppose, the height of

Ratan is 178 cm, then he is not a tall person according to the crisp set, albeit his height is very close to the 180 cm. Here, a person may either tall or not tall. No such thing in between tall or not tall. But, in reality, should we recognize a person as a tall person if and only if his height is exactly 180 cm or more than that? Indeed, the answer is no. One should not go for the crisp set representation in this case. However fuzzy set can provide the "in between" information by using the membership function. Suppose, the height of Ratan is 175 cm, then tallness of Ratan being more or less fulfilled. But, how much it fulfils the tallness? This can be found out by using the membership function. Here, the membership grade of tallness of Ratan is 0.8. (Fig. 2)

3.4 Fuzzy relation:

Relationship between events are specified by relation. Fuzzy relation is different from the crisp relation. In the following, crisp relation and fuzzy relation are illustrated with examples.

$$A = \{X_1, X_2, X_3\}$$

$$B = \{Y_1, Y_2, Y_3\}$$

Where, A and B are two universal set. X denotes the size of the flower and Y represents the market price.

X_1= Small , X_2= Medium, X_3=Large and Y_1=Low, Y_2=Medium, Y_3=High.

The crisp relation $X \rightarrow Y$ is defined as follows (Table 1):

Table 1: The crisp relation between X and Y

	Low (Y_1)	**Medium (Y_2)**	**High (Y_3)**
Small (X_1)	1	0	0
Medium (X_2)	0	1	0
Large (X_3)	0	0	1

The above table represents the crisp relationship between A and B. Here, 1 represents an association and 0 represents null assosciation. In linguistic term, one can write the above table as:

I. If the size of flower is small then the market price will be low.

II. If the size of flower is medium then the market price will be medium.

III. If the size of flower is large then the market price will be high.

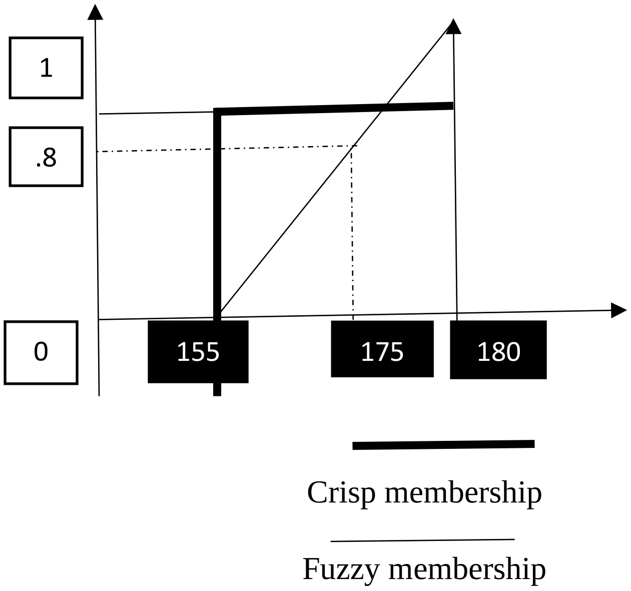


Figure 2: Fuzzy and crisp representation of A set

The fuzzy relation $X \rightarrow Y$ is defined as follows (Table 2):

Table 2: The fuzzy relation between X and Y

	Low (Y_1)	**Medium (Y_2)**	**High (Y_3)**
Small (X_1)	1	0.3	0
Medium (X_2)	0.3	1	0.3
Large (X_3)	0	0.3	1

The above table represents the fuzzy relationship between A and B. Here, the values represent the membership grade. In linguistic term, one may write the above table as-

I. If the size of flower is small then the market price will be low with the membership grade 1 and medium with 0.3 membership grade.

II. If the size of flower is medium then the market price will be medium with the membership grade 1, low with 0.3 membership grade and high with 0.3 membership grade

III. If the size of flower is Large then the market price will be large with the membership grade 1, medium with 0.3 membership grade.

Definition: A relation R which is a mapping from the cartesian space $X \times Y$ to the closed interval [0,1]- This relaion is called as fuzzy relation, where X and $Y$are two crisp set.

Membership function which represents the strength of mapping is denoted by:-

$$F_A(x, y) = \min\,(F_A(x), F_B(y)\,) \tag{5}$$

Where, A and B are two fuzzy sets.

Fuzzy cartesian product:

$$A = 0.5/x_1 + 0.6/x_2 + 0.2/x_3$$

$$B = 0.3/y_1 + 0.9/y_2$$

$$A \times B = R = \begin{bmatrix} 0.3 & 0.5 \\ 0.3 & 0.6 \\ 0.2 & 0.2 \end{bmatrix}$$

The above matrix is formed by taking the minimum value of each pair of elements.

3.5 *Max-Min Composition operator*:

There are many ways to combine the relations. Union or intersection operator can be utilized to combine the relation. Max-min composition operation is a kind of operator which is used for the combination of relation.

Let, two fuzzy relation A and B; then max-min composition between A and B is defined as:

$R = AoB = \left[(x, z), max\left\{min\left(F_A(x, y), F_B(y, z)\right)\right\}\right]$ for all $x \leftarrow X$,$y \leftarrow Y,$ $z \leftarrow Z$ (6)

A is defined on $X \times Y$.

B is defined on $Y \times Z$.

Example:

Let, $A = \begin{bmatrix} 0.7 & 0.6 \\ 0.2 & 0.5 \\ 0.3 & 0.6 \end{bmatrix}$ and $B = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.1 \end{bmatrix}$

$$R = AoB = \begin{bmatrix} 0.7 & 0.4 \\ 0.4 & 0.2 \\ 0.4 & 0.3 \end{bmatrix}$$

1st term of AoB,

$$r_{11} = max[min(a_{11}, b_{11}), min(a_{12}, b_{21})]$$

$$= max[min(0.7,0.8), min(0.6,0.4)]$$

$$= max[0.7,0.4]$$

$$= 0.7$$

Last term of AoB,

$$r_{32} = max[min(a_{31}, b_{12}), min(a_{32}, b_{22})]$$

$$= max[min(0.3,0.4), min(0.6,0.1)]$$

$$= max[0.3, 0.1]$$

$$= 0.3$$

3.6 Operations on fuzzy sets:

The basic operations on fuzzy set are union, intersection and complement. These operations are also the basic operation of the classical set theory. Though the name of the operations are same for both the classical and fuzzy set, but the processes are different.

Union: Let, two fuzzy sets are A and B and their membership functions are $m_A(x)$ and $m_B(x)$ respectively. Then, the union of A and B is the maximum value between the $m_A(x)$ and $m_B(x)$. If, the union is denoted by C, then the membership function of C is represented as:

$$m_C(x)=m_{A \cup B}(x) = max[m_A(x) , m_B(x)]$$
(7)

Intersection: The intersection of A and B is the minimum value between the $m_A(x)$ and $m_B(x)$. If, the intersection is denoted by C, then the membership function of C is represented as:

$$m_C(x)=m_{A \cap B}(x) = min[m_A(x) , m_B(x)] \quad (8)$$

Complement: Let, A is a fuzzy set and its membership function is $m_A(x)$. Then, the complement of A is denoted as $\bar{A}$ with its membership function is,

$$m_{\bar{A}}(x) = 1 - m_A(x) \quad (9)$$

4. Fuzzy time series:

Fuzzy time series is the time series with the fuzzy data which is based on fuzzy set theory. Song and Chissom (1993) first proposed fuzzy time series models employing max-min composition operation. They developed a step by step procedure to get the forecast and assessed the proposed model and verified the models' robustness property.

4.1 Definition: Let, fuzzy sets $f_i(t)$ are defined on $Y(t)$ and $F(t)$ is the collection of $f_i(t)$. Then, $F(t)$ is known as a fuzzy time series on $Y(t)$.

4.2 Fuzzification: It is the process of conversion from crisp data into fuzzy data.

4.3 Fuzzy logical relationship: If a fuzzy set A_1 is caused only by A_2, then fuzzy logical relationship is denoted by $\mathrm{A}_2 \to \mathrm{A}_1$.

4.4 Fuzzy logical relationship group: If some fuzzy logical relationship are $\mathrm{A}_2 \to A_1, A_2 \to A_3, A_2 \to A_2$; then fuzzy logical relationship group is denoted by $A_2 \to A_1, A_2, A_3$.

4.5 Defuzzification: Defuzzification is the process of conversion of fuzzified data into the crisp format. Actually, defuzzification is the counterpart of the fuzzification process. In literature, there are many methods for the fuzzification process. Some methods of defuzzification are described in the following:

4.5.1 Centroid Method**:** Centroid method is a method of weighted average in which it determines the centroid value of the sets. It is also known as centre of area method. The mathematical formula of this method is-

$$Y = \frac{\sum_{i=1}^{n} \omega_F(x_i)x_i}{\sum_{i=1}^{n} \omega_F(x_i)}$$

(10)

Where, Y is the crisp output.

$\omega_F(x_i)$ is the fuzzy output value or value of the membership function of x_i

x_i is the value of the element on the $x\ axis_i$ in the fuzzy set F.

4.5.2 Maximum membership method: This method gives the crisp output value which is equal to the value of x associated with the maximum value of membership. It can be expressed as:

$\omega_F(x_\$) \geq \omega_F(x)$ for all $x \in F$;

Where, $x_{\$}$ is the value associated with the maximum value of membership.

4.5.3 Average maximum membership method: It is like the maximum membership method but the only difference is that it may include more point other than the maximum point. It may include points which belong to some range.

5. General Steps involved in Fuzzy Time Series Forecasting Model

Most of the fuzzy time series forecasting model follow the following steps in forecasting process.

Step-1: Fixing the universe of discourse which is defined as- $U = [U_{min}$ -U_1, U_{max} -$U_2]$, where U_{min} and U_{max} are minimum and maximum value of the data and U_1 and U_2 are two any two positive values which are selected by the modeler properly. Define the proper universe of discourse to accommodate whole time series data.

Step 2: Division of the universe of discourse or define the intervals.

Step 3: Define fuzzy sets on the universe of discourse.

Step 4: Fuzzify the data which are based on the universe of discourse and corresponding fuzzy set defined in step-2 and step-3.

Step 5: Make the fuzzy logical relationship (FLR).

Step 6: Prepare the fuzzy logical relational groups.

Step 7: Forecast the time series data.

Step 8: Defuzzification of the forecasted fuzzified outputs.

6. Bibliography:

Saha, A., Singh, K. N., Ray, M., Kumar, S. and Rathod, S. (2019). A New Approach for Spatio-Temporal Modelling and Forecasting based on Fuzzy Techniques in conjunction with K- means clustering. *Journal of the Indian Society of Agricultural Statistics,* 73(2):111–120.

Song, Q. and Chissom, B.S. (1993a). Forecasting enrollments with fuzzy time series-part I. *Fuzzy Sets and Sys*., 54:1-10.

Song, Q. and Chissom, B.S. (1993b). Fuzzy time series and its models. *Fuzzy Sets and Sys*., 54: 269-277.

Zadeh, L.A. (1965). *Fuzzy sets,Inform. and Control*, 8:338-353.

Panel Data Regression Model

Ravindra Singh Shekhawat, Bishal Gurung and Achal Lama
ICAR-IASRI, New Delhi
Ravindra.Shekhawat@icar.gov.in

What is Panel Data

A data set containing observations on multiple phenomena observed over multiple time periods is called panel data.

Panel data gives more variability, more information, more efficiency and more degrees of freedom compared to the time series data or cross-section data.

The regression models based on such panel data are known as panel data regression models

1. Introduction

There are different types of data that are generally available for empirical analysis, namely, time series, cross section, and panel. A data set containing observations on a single phenomenon observed over multiple time periods is called time series (e.g., GDP for several quarters or years).

In time series data, both the *values* and the *ordering* of the data points have meaning. In cross-section data, values of one or more variables are collected for several sample units, or entities, at the same point of time (e.g., crime rates for 50 districts in the India for a given year).

A data set containing observations on multiple phenomena observed over multiple time periods is called panel data. In panel data the same cross-sectional units (say a family or a firm or a state) is surveyed over time. In short, panel data have space as well as time dimensions.

Let us consider a data set on eggs produced and their prices for 50 districts in India for years 2015 and 2016. For any given year, the data on eggs and their prices represent a cross-sectional sample. For any given district, there are two-time series observations on eggs and their prices. Thus, we have in all $(50 \times 2) = 100$ (panel) observations on eggs produced and their prices.

Some other examples:

- Data on yield of rice in 42 villages from 2013 to 2017, for 210 observations total.
- Data on crime rate in 17 Indian states, each state is observed in 6 years, for a total of 102 observations.
- Data on income of 1000 individuals, in four different months, for 4000 observations total.

There are other names for panel data, such as pooled data (pooling of time series and cross-sectional observations), combination of time series and cross-section data, micro panel data, longitudinal data (a study over time of a variable or group of subjects), cohort analysis (e.g., following the career path of 2004 graduates of an agricultural university). The regression models based on such panel data are known as panel data regression models.

Advantages of panel data

1. Since the panel data relate to individuals, firms, states, countries, etc., over time, there is bound to be heterogeneity in these units. The techniques of panel data estimation can take such heterogeneity explicitly into account by allowing for individual specific variables.
2. By combining time series of cross section observations, panel data give "more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency".
3. By studying the repeated cross section of observations, panel data are better suited to study the dynamics of change. Spells of unemployment, job turnover, and labour mobility are better suited with panel data.
4. Panel data can better detect and measure effects that simply cannot be observed in pure cross section or time series data.
5. Panel data enables us to study more complicated behavioral models. For example, phenomena such as economies of scale and technological change can be better handled by panel data than by pure cross section or time series data.
6. By making data available for several thousand units, panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregates.
7. More accurate inference of model parameters. Panel data usually contain more degrees of freedom and more sample variability than cross-sectional data which may be viewed as a panel with $T = 1$ (T is the number of time series), or time series data which is a panel with $N = 1$(N is the number of cross section), hence improving the efficiency of econometric estimates.

Panel Data: An illustrative example

Table 1.1

Obs	Y	X2	X3
_GE-1935	33.10000	1170.600	97.80000
_GE-1936	45.00000	2015.800	104.4000
_GE-1937	77.20000	2803.300	118.0000
_GE-1938	44.60000	2039.700	156.2000
_GE-1939	48.10000	2256.200	172.6000
_GE-1940	74.40000	2132.200	186.6000
_GE-1941	113.0000	1834.100	220.9000
_GE-1942	91.90000	1588.000	287.8000
_GE-1943	61.30000	1749.400	319.9000
_GE-1944	56.80000	1687.200	321.3000
_GE-1945	93.60000	2007.700	319.6000
_GE-1946	159.9000	2208.300	346.0000
_GE-1947	147.2000	1656.700	456.4000
_GE-1948	146.3000	1604.400	543.4000
_GE-1949	98.30000	1431.800	618.3000
_GE-1950	93.50000	1610.500	647.4000
_GE-1951	135.2000	1819.400	671.3000
_GE-1952	157.3000	2079.700	726.1000
_GE-1953	179.5000	2371.600	800.3000
_GE-1954	189.6000	2759.900	888.9000
_GM-1935	317.6000	3078.500	2.800000
_GM-1936	391.8000	4661.700	52.60000
_GM-1937	410.6000	5387.100	156.9000
_GM-1938	257.7000	2792.200	209.2000
_GM-1939	330.8000	4313.200	203.4000
_GM-1940	461.2000	4643.900	207.2000
_GM-1941	512.0000	4551.200	255.2000
_GM-1942	448.0000	3244.100	303.7000
_GM-1943	499.6000	4053.700	264.1000
_GM-1944	547.5000	4379.300	201.6000

_GM-1945	561.2000	4840.900	265.0000	
_GM-1946	688.1000	4900.000	402.2000	
_GM-1947	568.9000	3526.500	761.5000	
_GM-1948	529.2000	3245.700	922.4000	
_GM-1949	555.1000	3700.200	1020.100	
_GM-1950	642.9000	3755.600	1099.000	
_GM-1951	755.9000	4833.000	1207.700	
_GM-1952	891.2000	4924.900	1430.500	
_GM-1953	1304.400	6241.700	1777.300	
_GM-1954	1486.700	5593.600	2226.300	
_US-1935	209.9000	1362.400	53.80000	
_US-1936	355.3000	1807.100	50.50000	
_US-1937	469.9000	2673.300	118.1000	
_US-1938	262.3000	1801.900	260.2000	
_US-1939	230.4000	1957.300	312.7000	
_US-1940	361.6000	2202.900	254.2000	
_US-1941	472.8000	2380.500	261.4000	
_US-1942	445.6000	2168.600	298.7000	
_US-1943	361.6000	1985.100	301.8000	
_US-1944	288.2000	1813.900	279.1000	
_US-1945	258.7000	1850.200	213.8000	
_US-1946	420.3000	2067.700	232.6000	
_US-1947	420.5000	1796.700	264.8000	
_US-1948	494.5000	1625.800	306.9000	
_US-1949	405.1000	1667.000	351.1000	
_US-1950	418.8000	1677.400	357.8000	
_US-1951	588.2000	2289.500	341.1000	
_US-1952	645.2000	2159.400	444.2000	
_US-1953	641.0000	2031.300	623.6000	
_US-1954	459.3000	2115.500	669.7000	
_WEST-1935		12.93000	191.5000	1.800000
_WEST-1936		25.90000	516.0000	0.800000
_WEST-1937		35.05000	729.0000	7.400000
_WEST-1938		22.89000	560.4000	18.10000
_WEST-1939		18.84000	519.9000	23.50000
_WEST-1940		28.57000	628.5000	26.50000
_WEST-1941		48.51000	537.1000	36.20000
_WEST-1942		43.34000	561.2000	60.80000
_WEST-1943		37.02000	617.2000	84.40000
_WEST-1944		37.81000	626.7000	91.20000
_WEST-1945		39.27000	737.2000	92.40000
_WEST-1946		53.46000	760.5000	86.00000
_WEST-1947		55.56000	581.4000	111.1000
_WEST-1948		49.56000	662.3000	130.6000
_WEST-1949		32.04000	583.8000	141.8000
_WEST-1950		32.24000	635.2000	136.7000
_WEST-1951		54.38000	732.8000	129.7000
_WEST-1952		71.78000	864.1000	145.5000
_WEST-1953		90.08000	1193.500	174.8000
_WEST-1954		68.60000	1188.900	213.5000

Consider the data given in table 1.1, which are taken from a famous study of investment theory proposed by Y. Grunfeld (1958).

Grunfeld was interested in finding out how real gross investment (Y) depends on the real value of the firm (X_2) and real capital stock (X_3). We have data on four companies, General electric (GE), General Motor (GM), U.S. Steel (US), and Westinghouse. Data for each company on the preceding three variables are available for the period 1935-54. Thus, there are four cross-sectional units and 20 time periods. In all, therefore, we have 80 observations. A prior, Y is expected to be positively related to X_2 and X_3.

Pooling, or combining, all the 80 observations, the Grunfeld investment function can be written as:

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it}$$

$$i = 1, 2, 3, 4 \qquad (1.2.1)$$

$$t = 1, 2, \ldots, 20$$

where *i* stands for the *i*th cross-sectional unit and *t* for the *t*th time period and it is assumed that the X's are nonstochastic and that the error term follows the classical assumptions, namely, $E(\upsilon_{it}) \sim N(0, \sigma^2)$.

How do we estimate (1.2.1)? The answer follows.

2. Estimation of panel data regression models:

2.1The fixed effects approach.

Estimation of (1.2.1) depends on the assumptions we make about the intercept, the slope coefficients, and the error term. There are several possibilities:

1. Assume that the intercept and slope coefficients are constant across time and space and the error term captures differences over time and individuals.
2. The slope coefficients are constant but the intercept varies over individuals.
3. The slope coefficients are constant but the intercept varies over individuals and time.
4. All coefficients (the intercept as well as slope coefficients) vary over individuals.
5. The intercept as well as slope coefficients vay over individuals and time.

2.1.1 All coefficients constant across time and individuals

The simplest, and possibly naïve approach is to disregard the space and time dimensions of the pooled data and just estimate the usual OLS regression. That is, stack the 20 observations for each company one on top of the other, thus giving in all 80 observations for each of the variables in the model.

The OLS results are as follows

$$\hat{Y} = -63.3041 + 0.1101X_2 + 0.3034X_3$$

$$\begin{aligned} &se = (29.6124) \quad (0.0137) \quad (0.0493) \\ &t = (2.1376) \quad (8.0188) \quad (6.1545) \qquad\qquad (1.3.1) \\ &R^2 = 0.7565 \qquad \text{Durbin-Watson} = 0.2187 \\ &\qquad\qquad\qquad n = 80 \quad df = 77 \end{aligned}$$

Here all the coefficients are individually statistically significant and the R^2 value is reasonably high. But the only problem seems to be the estimated Durbin-Watson statistic which is quite low, suggesting that perhaps there is autocorrelation in the data.

The estimated model assumes that the intercept value of GE, GM, US, and Westinghouse are the same. It also assumes that the slope coefficients of two X variables are all identical for all the four firms. Obviously, these are very restricted assumptions. Therefore despite its simplicity the pooled regression may distort the true picture of the relationship between Y and X's across the four companies. What we need to do is find some way to take into account the specific nature of the four companies. How this can be done is explained next.

2.1.2 The slope coefficients are constant but the intercept varies over individuals: The Fixed Effects or Least-Squares Dummy Variables (LSDV) Regression Model

One way to take into account the individuality of each company or each cross-sectional unit is to let the intercept vary for each company but still assume that the slope coefficients are constant across firms. We write the model as:

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it} \qquad (1.3.2)$$

The difference in the intercept may be due managerial style or managerial philosophy.

The model (1.3.2)is known as the fixed effects (regression) model (FEM). The term "fixed effects" is due to the fact that, although the intercept may differ across individuals, each individual's intercept does not vary over time; that is, it is time invariant.

How do we actually allow for the (fixed effect) intercept to vary between companies? We can easily do that by the dummy variable technique. Therefore we write the model as

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it} \qquad (1.3.3)$$

where $D_{2i} = 1$ if the observation belongs to GM, 0 otherwise; $D_{3i} = 1$ if the observation belongs to US, 0 otherwise; and $D_{4i} = 1$ if the observation belongs to WEST, 0 otherwise. Here α_1 represents the intercept of GE and $\alpha_2, \alpha_3, \text{and}\, \alpha_4$, the differential intercept coefficients, tell by how much the intercepts of GM, US, and WEST differ from the intercept of GE.

Since we are using dummies to estimate the fixed effects, the model is also known as the least-squares dummy variable (LSDV) model.

The results based on (1.3.3) are as follows:

$$\hat{Y}_{it} = -245.7924 + 161.5722 D_{2i} + 339.6328 D_{3i} + 186.5666 D_{4i} + 0.1079 X_{2i} + 0.3461 X_{3i}$$

se = (35.8112) (46.4563) (23.9863) (31.5068) (0.0175) (0.0266)

t = (6.8635) (3.4779) (14.1594) (5.9214) (6.1653) (12.9821)

$R^2 = 0.9345$ d =1.1076 df = 74 (1.3.4)

In (1.3.4) all the estimated coefficients are individually highly significant and the intercept values of the four companies are statistically different. The differences in the intercepts may be due to unique features of each company, such as differences in management style or managerial talent.

Judged by the statistical significance of the estimated coefficients, and the fact that the R^2 value has increased substantially we can conclude that (1.3.4) is better than (1.3.1). The Durbin-Watson *d* value is much higher, suggesting that model (1.3.1) was mis-specified.

We can also provide a formal test of the two models. In relation to (1.3.4), model (1.3.1) is a restricted model in that it imposes a common intercept on all the companies. Therefore, we can use the restricted F test. Using the formula we get

$$F = \frac{(R^2_{UR} - R^2_{R})/3}{(1 - R^2_{UR})/74} = 66.9980 \qquad (1.3.5)$$

where the restricted value is from (1.3.1) and the unrestricted is from (1.3.4).

Clearly, the F value of 66.998 is highly significant and, therefore, the restricted regression (1.3.1) seems to be invalid.

The Time Effect.

Just as we used the dummy variables to account for individual effect, we can allow for time effect in the sense that the Grunfeld investment function shifts over time. For such a situation we introduce time dummies, one for each year.

$$Y_{it} = \lambda_0 + \lambda_1 D35 + \lambda_2 D36 + \ldots + \lambda_{19} D53 + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it} \qquad (1.3.6)$$

From the regression results, we infer that none of the individual time dummies were individual statistically significant.
We have already seen that the individual company effects were statistically significant, but the individual year effects were not. Could it be that our model is mis-specified in that we have not taken into account both individual and time effects together? Let us consider this possibility.

2.1.3 Slope coefficients constant but the intercept varies over individual as well as time

To consider this possibility, we can combine (1.3.4) and the time effect model, as follows:

$$Y_{it} = \alpha_1 + \alpha_2 D_{GMi} + \alpha_3 D_{USi} + \alpha_4 D_{WESTi} + \lambda_0 + \lambda_1 DUM35 + \ldots + \lambda_{19} DUM53 + \ldots$$
$$+ \beta_2 X_{2i} + \beta_3 X_{3i} + \upsilon_{it} \qquad (1.3.7)$$

when we run this regression, we find the company dummies as well as the coefficients of the X are individually statistically significant, but none of the time dummies are. Essentially, we are back to (1.3.4).

2.1.4 All coefficients vary across individuals

Here we assume that the intercepts and the slope coefficients are different for all individual, or cross-section, units. This is to say that the investment functions of GE, GM, US, and WEST are all different. We can easily extend our LSDV model to take care of this situation. Here what we do is multiply each of the company dummies by each of the X variables.

That is, we estimate the following model:

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_{3it} + \gamma_1 (D_{2i} X_{2it}) + \gamma_2 (D_{2i} X_{3it}) +$$
$$\gamma_3 (D_{3i} X_{2it}) + \ldots + \upsilon_{it} \qquad (1.3.8)$$

The γ's are the differential slope coefficients, just as $\alpha_2, \alpha_3,$ and α_4 are the differential intercepts. If one or more of the γ coefficients are statistically significant, it will tell us that one or more slope coefficients are different from the base group. If all the differential intercept and all the differential slope coefficients are statistically significant, we can conclude that the investment functions are different for the four companies.

A caution on the use of the Fixed Effects, or LSDV, model.

Although easy to use, the LSDV model has some problems that need to be bourne in mind.
First, if you introduce too many dummy variables we will run up against the degrees of freedom problem.
Second, with so many variables in the model, there is always the possibility of multicollinearity, which might make precise estimation of one or more parameters difficult.
Third, suppose in the FEM if variables such as sex, color, and ethnicity, which are time invariant are also included, the LSDV approach may not be able to identify the impact of such time-invariant variables.

Estimation of panel data regression models:

2.2 The random effects approach

Although fixed effects or LSDV model can be expensive in terms of degrees of freedom if we have several cross-sectional units.
If the dummy variables do in fact represent a lack of knowledge about the (true) model, why not express this ignorance through the disturbance term υ_{it} ? This is precisely the approach suggested by the proponents of the so called error components model (ECM) or random effects model (REM).
The basic idea is to start with (1.3.2):

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it} \qquad (1.4.1)$$

Instead of treating β_{1i} as fixed, we assume that it is a random variable with a mean value of β_1. And the intercept value for an individual company can be expressed as

$$\beta_{1i} = \beta_1 + \varepsilon_i \qquad i = 1, 2, \ldots, N \qquad (1.4.2)$$

where ε_i is a random error term with a mean value of zero and variance σ_ε^2.
What we are essentially saying is that the four firms included in our sample are a drawing from a much larger universe of such companies and that they have a common mean value for the intercept and the individual differences in the intercept values of each company are reflected in the error term.

Substituting (1.4.2) into (1.4.1), we get:

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + \upsilon_{it}$$

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \omega_{it} \qquad (1.4.3)$$

where $$\omega_{it} = \varepsilon_i + \upsilon_{it} \qquad (1.4.4)$$

The composite error term consist of two components, the cross-section, or individual-specific, error component and the combined time series and cross-section error component.
The usual assumption made by ECM are that

$$\begin{aligned}
&\varepsilon_i \sim N(0,\sigma_\varepsilon^2) \\
&\upsilon_{it} \sim N(0,\sigma_\upsilon^2) \\
&E(\varepsilon_i \upsilon_{it}) = 0 \qquad E(\varepsilon_i \varepsilon_j) = 0 \qquad (i \neq j) \\
&E(\upsilon_{it} \upsilon_{is}) = E(\upsilon_{it} \upsilon_{jt}) = E(\upsilon_{it} \upsilon_{js}) = 0 \qquad (i \neq j; t \neq s)
\end{aligned} \qquad (1.4.5)$$

that is, the individual error components are not correlated with each other and are not autocorrelated across both cross-section and time series units.
As a result it follows that

$$E(\omega_{it}) = 0 \qquad (1.4.6)$$

$$\text{var}(\omega_{it}) = \sigma_\varepsilon^2 + \sigma_\upsilon^2 \qquad (1.4.7)$$

As (1.4.7) shows, the error term ω_{it} is homoscedastic. However, it can be shown that ω_{it} and ω_{is} are correlated; that is, the error terms of a given cross-sectional unit at two different points in time are correlated. The correlation coefficient is as follows:

$$\text{corr}(\omega_{it}, \omega_{is}) = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\upsilon^2} \qquad (1.4.8)$$

If we do not take this correlation structure into account, and estimate (1.4.3) by OLS, the resulting estimators will be inefficient. The most appropriate method here is the method of Generalized least squares (GLS).

The result of the model is given below

$$\begin{aligned}
&\hat{Y} = -73.03 + 0.1076X_2 + 0.3457X_3 \\
&se = (83.9495)\,(0.0168) \quad (0.0168) \\
&t = (0.8699) \quad (6.4016) \quad (13.0235) \\
&\qquad\qquad R^2 = 0.9323
\end{aligned}$$

3. Fixed Effects (LSDV) versus Random Effects Model

The challenge facing a researcher is: which model is better, FEM or ECM? The answer to this question hinges around the assumption one makes about the likely correlation between the individual, or cross-section specific, error components and the X regressors.
If it is assumed that the error component and the X's are uncorrelated, ECM may be appropriate, whereas if they are correlated, FEM may be appropriate.
Keeping this fundamental difference in the two approaches in mind, what more can we say about the choice between FEM and ECM?

The answer may be:

1. If T (the number of time series data) is large and N (the number of cross-sectional units) is small, there is likely to be little difference in the values of the parameters estimated by FEM and ECM. Hence the choice here is based on computational convenience. On this score, FEM may be preferable.

2. When N is large and T is small, the estimates obtained by the two methods can differ significantly. Recall that in ECM $\beta_{1i} = \beta_1 + \varepsilon_i$, but in FEM we treat β_{1i} as fixed and non-random.
3. If the individual error component ε_i and one or more regressors are correlated, then the ECM estimators are biased, whereas those obtained from FEM are unbiased.
4. If N is large and T is small, and if the assumptions underlying ECM hold, ECM estimators are more efficient than FEM estimators.

Is there a formal test that will help us to choose between FEM and ECM? Yes, a test was developed by Hausman in 1978. The null hypothesis underlying the Hausman test is that the FEM and ECM estimators do not differ substantially. The test statistic developed by Hausman has an asymptotic chi-square distribution. If the null hypothesis is rejected, the conclusion is that ECM is not appropriate and that we may be better off using FEM, in which case statistical inferences will be conditional on the error component in the sample.

Conclusion

1. Panel data, by blending the inter-individual differences and intra-individual dynamics have advantages over cross-sectional or time-series data.
2. Greater capacity for capturing the complexity of human behavior than a single cross-section or time series data.
3. More accurate inference of model parameters can be obtained through panel data.
4. Panel data usually contain more degrees of freedom and more sample variability than cross-sectional data or time series.
5. Controlling the impact of omitted variables, i.e., reduces omitted variable bias.
6. Panel data helps in uncovering dynamic relationships.
7. Panel regression models are based on panel data. Panel data consist of observations on the same cross-sectional, or individual, units over several time periods.
8. There are several advantage to using panel data. First, they increase the sample size considerably. Second, by studying repeated cross-section observations, panel data are better suited to study the dynamics of changes. Third, Panel data enable us to study more complicated behavioral models.
9. Despite their substantial advantages, panel data pose several estimation and inference problems. Since such data involve both cross-section and time dimensions, problem that plague cross-sectional data (eg., heteroscedasticity) and time series data (eg., autocorrelation) need to be addressed. There are some additional problems, such as cross-correlation in individual units at the same point of time.
10. There are several estimation techniques to address one or more of these problems. The two most prominent are (1) the fixed effect model (FEM) and (2) the random effect model (REM) or error component model (ECM).
11. In FEM, the intercept in the regression model is allowed to differ among individuals, or cross-sectional, unit may have some special characteristics of its own. To take in to account the differing intercepts, one can use dummy variables. The FEM using dummy variables is known as the least-squares dummy variable model (LSDV). FEM is appropriate in a situation where the individual-specific intercept may be correlated with one or more repressors. A disadvantage of LSDV is that it consumes a lot of degree of freedom when the number of the cross sectional units, N, is very large, in

which case we will have to introduce N dummies (but suppress the common intercept term).

12. An alternative of FEM is ECM. In ECM it is assumed that the intercept of an individual unit is a random drawing from a much larger population with a constant mean value. The individual intercept is then expressed as a deviation from this constant mean value. One advantage of ECM over FEM is that it is economical in degree of freedom, as we do not have to estimate N cross-sectional intercept. We need only to estimate the mean value of the intercept and its variance. ECM is appropriate in situation where the (random) intercept of each cross-sectional unit is uncorrelated with repressors.
13. The Hausman test can be used to decide between FEM and ECM.

Bibliography

Ahn, S.C. and P. Schmidt (1995), " Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, 68, 5-27.

Arellano, M., (2003), *Panel Data Econometrics*, Oxford: Oxford University Press.

Balestra, P. and M. Nerlove (1966), "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas", *Econometrica*, 34, 585-612.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, Second edition, New York: Wiley.

Chamberlain, G. (1984), "Panel Data", in *Handbook of Econometrics* Vol II, ed. by Z. Griliches and M. Intriligator, pp. 1247-1318. Amsterdam: North Holland.

Frees, E. (2004). Longitudinal and Panel Data, Cambridge University Press.

Grunfeld, Y. (1958), "The Determinants of Corporate Investment", unpublished Ph.D. thesis, Department of Economics, University of Chicago.

Hausman, J.A. (1978), "Specification Tests in Econometrics", *Econometrica*, 46. 1251-71.

Hsiao, C., (1986) "*Analysis of Panel Data*, Econometric Society monographs No. 11, New York: Cambridge University Press.

Hsiao, C. (2003), *Analysis of Panel Data*, 2nd edition, Cambridge: Cambridge University Press (Econometric Society monograph no. 34).

Rao, C.R., (1973), Linear Statistical Inference and Its Applications, 2nd ed., New York: Wiley.

Modelling and Forecasting of Volatile Time-Series Data

Achal Lama, Bishal Gurung and R S Shekhawat
ICAR-IASRI, New Delhi
achal.lama@icar.gov.in

1. Introduction

In agriculture, one of the most prominent sectors in India, data are usually collected over time. Linear Gaussian models are not able to describe changing conditional variance, which is present in many such real data sets. To handle such a situation, Engle (1982) introduced the Autoregressive conditional heteroscedastic (ARCH) models in which significant presence of autocorrelation of squared residual series is considered.

The ARCH (q) model for series $\{\varepsilon_t\}$ is defined by specifying the conditional distribution of ε_t given information available up to time *t-1*. The process $\{\varepsilon_t\}$ is ARCH (q), if the conditional distribution of $\{\varepsilon_t\}$ given available information ψ_{t-1} is

$$\varepsilon_t|\psi_{t-1} \sim N(0, h_t) \tag{1.1}$$

and

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 \tag{1.2}$$

where $a_0 > 0, a_i \geq 0$ for all i and $\sum_{i=1}^{q} a_i < 1$

Ghosh and Prajneshu (2003) have applied the AR(p)-ARCH(q) model to study the volatility present in onion price data. The fitted model provided a significantly good description of underlying mechanism in terms of significant ARCH parameters and changing forecast interval of hold-out-data. Ghosh *et al*. (2005, 2006) have also studied various aspects of the family of mixtures of ARCH models. The AR-ARCH model has also been used as the basic “building blocks” for Markov switching and mixture models (Lanne and Saikkonen, 2003 and Wong and Li, 2001). Since Engle introduced ARCH model, various extensions of ARCH models have been proposed to model volatility. However, ARCH model has some drawbacks. Firstly, when the order of ARCH model is very large, estimation of a large number of parameters is required which is cumbersome. Secondly, the conditional variance of ARCH(q) model has the property that unconditional autocorrelation function (Acf) of squared residuals; if it exists, decays very rapidly compared to what is typically observed, unless maximum lag q is large.

To overcome these difficulties, Bollerslev (1986) proposed the Generalized ARCH (GARCH) model in which conditional variance is also a linear function of its own lags.

This model is also a weighted average of past squared residuals, but it has declining weights that never go completely to zero. It gives parsimonious models that are easy to estimate and, even in its simplest form, has proven surprisingly successful in predicting conditional variances. The GARCH model focuses on capturing the clustering of volatility in returns when the conditional variance at time *t* is modelled as a deterministic function of lagged values of conditional variances and squared returns, given by

$$\varepsilon_t = \xi_t h_t^{1/2} \qquad (1.3)$$

and

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j h_{t-j} \qquad (1.4)$$

where $\xi_t \sim IID(0,1),\ a_0 > 0, a_i \geq 0,\ i = 1,2,\dots,q.\ \ b_j \geq 0,\ \ j = 1,2,\dots,p$

Angelidis *et al*. (2004) evaluated the performance of GARCH models in modelling the daily Value-at-Risk (VaR) of perfectly distributed portfolios in five stock indices, using a number of distributional assumptions and sample sizes. However, the GARCH model cannot capture in a more appropriate way the main empirical properties often observed in volatile time-series data. To this end, Stochastic Volatility (SV) parametric nonlinear time-series model was proposed to capture time-varying volatility (Taylor, 1994). SV models the variance as an unobserved component that follows a particular stochastic process. This way the properties of SV models are more attractive and closer to the empirical properties observed in volatile time-series data. Although their estimation is more complicated, it gives parsimonious models. In this write up, an attempt is made to study the SV model along with its estimation procedure. The novelty of this work is that a new form of state-space modelling is proposed to estimate the volatility as well as the parameters of SV model. Powerful Kalman filtering technique has been employed to obtain the best linear predictors of unobserved volatility. Using the prediction error decomposition form of the likelihood, the Quasi-maximum likelihood is maximized with respect to the unknown parameters. The consistency of the estimated parameters is also validated. A posterior analysis of innovation based on log volatility and its lag value, when the information till time *t* is known, is carried out for validating the consistency of autoregressive coefficient of volatility process. The posterior analysis also gives the estimate of variance of log-volatility process which is compared with the one obtained by prediction error decomposition for validation.

As an illustration, this model is applied to describe the volatile All-India data of monthly export of spices during the period January, 2006 to January, 2012. Based on the residuals, the performance of the SV and GARCH models is assessed for modelling as well as forecasting. It is concluded that SV model performs relatively well for the data under consideration.

2. Some Preliminaries

In this section, Stochastic Volatility model along with its estimation procedure, the Kalman filtering, and ARCH-LM test is briefly described.

2.1 Stochastic Volatility (SV) Model

Time-series data of some agricultural commodities show some statistical properties, viz. Leptokurtic distributions, Volatility clustering and ARCH effect, meaning the squared residuals exhibit serial correlation whereas little or no serial dependence can be detected in the residual series itself. Since volatility evolves over time, modelling the volatility plays an important role in both parameter estimation and the accuracy in interval forecast. The SV model understands the time-varying variance as a stochastic process. It is also able to represent excess kurtosis as well as autocorrelations of squares.

Consider the univariate discrete time SV model (Taylor, 1994)

$$y_t = \varepsilon_t \sigma_t \ , \qquad \text{t = 1, . . . , T,} \qquad (2.1.1)$$

where y_t are observations, ε_t is a white noise process with unit variance and σ_t is the corresponding volatility. $log\sigma_t^2$ follows an AR (1) process with Gaussian noise and is unobserved but can be estimated using the observations. Following the convention usually considered in literature we write $h_t^* = log\sigma_t^2$. So, eq. (2.1.1) can be written as

$$y_t = \varepsilon_t exp\ (h_t^*/2) \qquad (2.1.2)$$

and

$$h_{t+1}^* = \alpha + \varphi h_t^* + \eta_t, \quad \eta_t \sim IID(0, \sigma_\eta^2), \quad |\varphi| < 1 \qquad (2.1.3)$$

where $|\varphi| < 1$ implies stationarity of h_t^*. The parameter φ measures the persistence of shocks to the volatility. When φ is close to *1* and σ_η^2 is close to *0*, the evolution of volatility over time is very smooth. The variance of the log-volatility process, σ_η^2 measures the uncertainty about future volatility.

Now Eq. (2.1.3) can be written as

$$(h_{t+1}^* - \alpha^*) = \varphi(h_t^* - \alpha^*) + \eta_t$$

where $\alpha^* = \ \alpha/(1-\varphi)$

So, Eq. (2.1.2) becomes

$$y_t = exp((h_t^* - \alpha^*)/2)\, exp(\alpha^*/2)\, \varepsilon_t$$

which can be written as

$$y_t = \sigma_* exp(h_t/2)\, \varepsilon_t$$

where $h_t = h_t^* - \alpha^*, \sigma_* = exp(\alpha^*/2)$

So the SV model can be rewritten as

$$y_t = \sigma_* exp(h_t/2)\, \varepsilon_t$$

$$h_{t+1} = \varphi h_t + \eta_t$$

The initial distribution of h_t denotes the unconditional distribution of the process *$\{h_t\}$* i.e. the unconditional distribution of h_t is commonly used as an initial condition. If ε_t has a finite variance, the variance of the observation can be written as

$$Var(y_t) = \sigma_*^2 \sigma_\varepsilon^2 exp(\sigma_h^2/2) \qquad (2.1.4)$$

If the fourth moment of ε_t exist, the kurtosis of y_t is given as $\kappa * exp\ (\sigma_h^2)$, where κ is the kurtosis of ε_t, so it can be seen that y_t exhibits more kurtosis than ε_t and all the odd moments are zero. The square of the coefficient of variation (CV) of σ_t^2 is used as a measure of the relative strength of the SV. This is given as

$$Var(\sigma_t^2)/[E(\sigma_t^2)]^2 = exp(\sigma_h^2) - 1 \qquad (2.1.5)$$

If η_t is assumed to be normal, the ACF of the absolute values of the observations raised to the power *c* is given by

$$\rho_\tau^{(c)} = \frac{E(|y_t|^c |y_{t-\tau}|^c) - \{E(|y_t|^c)\}^2}{E(|y_t|^{2c}) - \{E(|y_t|^c)\}^2} = \frac{exp\left(\frac{c^2}{4}\sigma_h^2 \rho_{h,\tau}\right) - 1}{\kappa_c exp\left(\frac{c^2}{4}\sigma_h^2\right) - 1} \quad , \tau \geq 1, c > -0.5, c \neq 0$$

(2.1.6)

where

$$\kappa_c = \frac{E(|y_t|^{2c})}{\{E(|y_t|^c)\}^2}$$

and $\rho_{h,\tau}, \tau = 0,1,2, \ldots$ denotes the ACF of h_t. Taylor (1986) derived this expression for *c* equal to one and two and ε_t normally distributed. When *c=2*, κ_c is the kurtosis and this is three for a normal distribution. More generally,

$$\kappa_c = \Gamma(c + 1/2)\Gamma(1/2)/\{\Gamma(c/2 + 1/2\}^2 \ , \ c \neq 0 \qquad (2.1.7)$$

The ACF, $\rho_\tau^{(c)}$, has the following features. First, if σ_h^2 is small and/or $\rho_{h,\tau}$ is close to one,

$$\rho_\tau^{(c)} \cong \rho_{h,\tau} \frac{exp\left(\frac{c^2}{4}\sigma_h^2\right) - 1}{\kappa_c exp\left(\frac{c^2}{4}\sigma_h^2\right) - 1} \quad , \tau \geq 1 \qquad (2.1.8)$$

Thus the shape of the ACF of h_t is approximately carried over to $\rho_\tau^{(c)}$except that it is multiplied by a factor of proportionality, which must be less than one for c positive as κ_c is greater than one. Although the series y_t is uncorrelated, it is not an independent sequence. The dynamics of the series appear in the squared residuals. Their autocorrelation function (acf), is given by

$$\rho_2(\tau) = \frac{exp\left(\sigma_h^2\rho_h(\tau)\right)-1}{\kappa_\varepsilon exp(\sigma_h^2)-1}, \qquad \tau \geq 1 \qquad (2.1.9)$$

where $\rho_h(\tau)$ is the autocorrelation of order τ of the underlying log-volatility.

Comparison with ARCH/GARCH models

The GARCH (1, 1) model has been applied extensively to volatile time-series data. In GARCH the variance is assumed to depend on the variance and squared observation in the previous time period. The motivation comes from forecasting; in an AR(1) model with independent disturbances, the optimal prediction of the next observation is a fraction of the current observation, and in ARCH it is a fraction of the current squared observation.

The specification of GARCH (1, 1) means that we can write

$$y_t^2 = \gamma + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2 + \upsilon_t = \gamma + (\alpha+\beta)y_{t-1}^2 + \upsilon_t - \beta\upsilon_{t-1} \qquad (2.1.10)$$

where $\upsilon_t = y_t^2 - \sigma_t^2$ is a martingale difference. Thus, y_t^2 has the form of an ARMA (1, 1) process and so the ACF can be evaluated. The GARCH (1, 1) displays similar properties to the SV model, particularly if φ is close to one. The main difference in the ACFs seems to show up most at lag one. Jacquier *et al*. (1994) presented a graph of the correlogram of the squared weekly returns of a portfolio on the New York Stock Exchange together with the ACFs of SV and GARCH (1, 1) models. It was seen that the ACF implied by the SV model was closer to the sample values.

The SV model displays excess kurtosis even if φ is zero since y_t is a mixture of distributions. The σ_η^2 parameters govern the degree of mixing independently of the degree of smoothness of the variance evolution. This is not the case with a GARCH model where the degree of kurtosis is tied to the roots of the variance equation. Moreover, the basic GARCH model does not allow for the kind of asymmetry captured by a SV model with contemporaneously correlated disturbances.

A good description of stochastic volatility models is given in Barndorff-Nielsen *et al*. (2002) and Broto and Ruiz (2004).

Linear state space form:

The dependence between ε_t and η_t allows the model to capture the kind of asymmetric behavior that is often found in commodity prices. Also the assumption of contemporary dependence explains conditional variance $h_{t+1} = log\sigma_{t+1}^2$ of price innovation in terms of past values. The logarithmic transformed observations, the $logy_t^2's$, can be used to construct a linear state space model. Moreover, the linear state space form can be modified so as to deal with asymmetric models. Even if η_t and $\varepsilon_t$are not mutually independent, the disturbances in the state space form are uncorrelated if the joint distribution of η_t and ε_t is symmetric. But, there is a loss of information by taking logarithm of the squared observations. Harvey and Shephard (1993) showed that this information may be recovered by conditioning on the signs of the observations denoted by s_t, a variable which takes the value +1(-1) when the observation is positive (negative).

Denote $E_+(E_-)$ as the expectation conditional on ε_t being positive (negative), and assign a similar interpretation to variance and covariance operators. The distribution of ξ_{t} is not affected by conditioning on the signs of the $\varepsilon_t's$, but, it should be kept in mind that $E(\eta_t|\varepsilon_t)$ is an odd function of ε_t. So,

$$\mu^* = E_+(\eta_t) = E_+[E(\eta_t|\varepsilon_t)] = -E_-(\eta_t), \qquad (2.1.11)$$

and

$$\gamma^* = Cov_+(\eta_t, \xi_{\mathrm{t}}) = E_+(\eta_t\xi_{\mathrm{t}}) - E_+(\eta_t)E(\xi_{\mathrm{t}}) = E_+(\eta_t\xi_{\mathrm{t}}) = -Cov_-(\eta_t, \xi_{\mathrm{t}})$$

(2.1.12)

because the expectation of $\xi_{\mathrm{t}}$is zero and

$$E_+(\eta_t\xi_{\mathrm{t}}) = E_+[E(\eta_t|\varepsilon_t)log\varepsilon_t] - \mu^*E(log\varepsilon_t) = -E_-(\eta_t\xi_{\mathrm{t}}) \qquad (2.1.13)$$

Finally,

$$Var_+(\eta_t) = E_+(\eta_t^2) - [E_+(\eta_t)]^2 = \sigma_\eta^2 - \mu^{*2} \qquad (2.1.14)$$

The linear state space form is now

$$\log(y_t^2) = \omega + h_t + \xi_t \qquad (2.1.15)$$

$$h_{t+1} = \varphi h_t + s_t\mu^* + \eta_t^* \qquad (2.1.16)$$

where

$$\omega = \log(\sigma_*^2) + E(\log(\varepsilon_{\mathrm{t}}^2)) \text{ , } h_t = \log(\sigma_t^2) \text{ and } \xi_{\mathrm{t}} = \log(\varepsilon_{\mathrm{t}}^2) - E(\log(\varepsilon_{\mathrm{t}}^2)).$$

$$\begin{pmatrix} \xi_{\mathrm{t}} \\ \eta_t^* \end{pmatrix} |s_t \sim ID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \gamma^* s_t \\ \gamma^* s_t & \sigma_\eta^2 - \mu^{*2} \end{pmatrix}\right)$$

The above model can be transformed to a more convenient form with uncorrelated measurement and transition equation errors. This is

$$\log(y_t^2) = \omega + h_t + \xi_t \qquad (2.1.17)$$

$$h_{t+1} = (\varphi - \gamma^* s_t/\sigma_\xi^2)h_t + s_t\{\mu^* + \gamma^*/\sigma_\xi^2\,(logy_t^2 - \omega)\} + \eta_t^+ \qquad (2.1.18)$$

$$\begin{pmatrix} \xi_t \\ \eta_t^+ \end{pmatrix} |s_t \sim ID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\eta^2 - \mu^{*2} - (\gamma^{*2}/\sigma_\xi^2) \end{pmatrix}\right)$$

The filtered estimate of the log volatility, written as $\hat{h}_{t+1/t}$ is given by

$$\hat{h}_{t+1/t} = \left(\varphi - \gamma^* s_t/\sigma_\xi^2\right)\left(\frac{\sigma_\xi^2}{p_{t/t-1}+\sigma_\xi^2}\right)\hat{h}_{t/t-1} + \frac{(logy_t^2-\omega)}{p_{t/t-1}+\sigma_\xi^2}\{\gamma^* s_t + \varphi p_{t/t-1}\} + s_t\mu^*$$

$$(2.1.19)$$

2.2 Kalman filter and state space model

State space modelling includes the State transition equation, eq. (2.2.1), which allows the state variable $\boldsymbol{\alpha}_t$ to change through time, and the Measurement equation, eq. (2.2.2), which relates the state variable to an observation Y_t.

$$\boldsymbol{\alpha}_{t+1} = \mathbf{F}_t\boldsymbol{\alpha}_t + \mathbf{G}_t\boldsymbol{\varepsilon}_t \qquad (2.2.1)$$

$$Y_t = \mathbf{H}_t'\boldsymbol{\alpha}_t + \nu_t \qquad (2.2.2)$$

It is assumed that $\{\boldsymbol{\varepsilon}_t\}$ of eq. (2.2.1) and $\{\nu_t\}$ of eq. (2.2.2) are independent, zero mean, Gaussian white noise process with

$$E[\nu_t\nu_t'] = R_t \quad \text{and} \quad E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \mathbf{Q}_t \qquad (2.2.3)$$

The Kalman filter (KF) is a recursive algorithm for sequentially updating the state vector given past information, $\boldsymbol{\psi}_t$.

Denote

$$\hat{\boldsymbol{\alpha}}_{t|t-1} = E\{\boldsymbol{\alpha}_t|\psi_{t-1}\} \text{ and } \hat{\boldsymbol{\alpha}}_{t|t} = E\{\boldsymbol{\alpha}_t|\psi_t\} \text{ for } t = 0,1,2,\ldots \qquad (2.2.4)$$

and assume $\hat{\boldsymbol{\alpha}}_{0|-1} = E\{\boldsymbol{\alpha}_0\}$ and $\boldsymbol{\Sigma}_{0|-1} = \mathbf{P}_0$. The state vector $\boldsymbol{\alpha}_t$ and its mean squared error $\boldsymbol{\Sigma}_t = E[(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)']$ are recursively estimated by:

$$\hat{\boldsymbol{\alpha}}_{t|t} = \hat{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t(\mathbf{H}'_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t + R_t)^{-1}(X_t - \mathbf{H}'_t\hat{\boldsymbol{\alpha}}_{t|t-1}) \qquad (2.2.5)$$

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t(\mathbf{H}'_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t + R_t)^{-1}\mathbf{H}'_t\boldsymbol{\Sigma}_{t|t-1} \qquad (2.2.6)$$

Using the recursive filter equations (2.2.1) and (2.2.2), we can obtain $\hat{\boldsymbol{\alpha}}_{t+1|t}$ as

$$\hat{\boldsymbol{\alpha}}_{t+1|t} = \mathbf{F}_t\hat{\boldsymbol{\alpha}}_{t|t} \qquad (2.2.7)$$

and

$$\boldsymbol{\Sigma}_{t+1|t} = \mathbf{F}_t\boldsymbol{\Sigma}_{t|t}\mathbf{F}'_t + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}'_t \qquad (2.2.8)$$

Eq. (2.2.7) can also be written as

$$\hat{\boldsymbol{\alpha}}_{\mathrm{t+1|t}} = \mathbf{F}_{\mathrm{t}}\hat{\boldsymbol{\alpha}}_{\mathrm{t|t-1}} + \mathbf{F}_{\mathrm{t}}\boldsymbol{\Sigma}_{\mathrm{t|t-1}}\mathbf{H}_{\mathrm{t}}(\mathbf{H'}_{\mathrm{t}}\boldsymbol{\Sigma}_{\mathrm{t|t-1}}\mathbf{H}_{\mathrm{t}} + \mathrm{R}_{\mathrm{t}})^{-1}(\mathrm{Y}_{\mathrm{t}} - \mathbf{H'}_{\mathrm{t}}\hat{\boldsymbol{\alpha}}_{\mathrm{t|t-1}}) \qquad (2.2.9)$$

which implies that the time update rules for each forecast of state are weighted average of the previous forecast $\hat{\boldsymbol{\alpha}}_{\mathrm{t|t-1}}$ and the forecast error $(\mathrm{Y}_{\mathrm{t}} - \mathbf{H'}_{\mathrm{t}}\hat{\boldsymbol{\alpha}}_{\mathrm{t|t-1}})$. After obtaining $\hat{\boldsymbol{\alpha}}_{\mathrm{t|t-1}}$, one may predict Y_{t} by the optimal predictor $\hat{\mathrm{Y}}_{\mathrm{t|t-1}}$, where

$$\hat{\mathrm{Y}}_{\mathrm{t|t-1}} = \mathbf{H'}_{\mathrm{t}}\hat{\boldsymbol{\alpha}}_{\mathrm{t|t-1}} \qquad (2.2.10)$$

and the conditional error variance due to predictor $\hat{\mathrm{Y}}_{\mathrm{t|t-1}}$ is

$$\mathbf{H'}_{\mathrm{t}}\boldsymbol{\Sigma}_{\mathrm{t|t-1}}\mathbf{H}_{\mathrm{t}} + \mathrm{R}_{\mathrm{t}} \qquad (2.2.11)$$

An excellent description of this methodology is given in Durbin and Koopman (2001).

2.3 Estimation of parameters

The parameters of the SV model are estimated using the KF technique in conjunction with Quasi-maximum likelihood (QML) principle. The QML estimator is based on maximizing the log-likelihood function even if the assumption of normality is violated (Harvey *et al.*, 1994). Andersen *et al.* (2001) also showed that the log-volatility process can be well approximated by a Normal distribution. The joint density of the observations can be expressed as

$$f(y_1, \dots, y_T; \theta) = \prod_{t=1}^{T} f(y_t|\psi_{\mathrm{t-1}}; \theta), \qquad (2.3.1)$$

where θ is the vector of unknown parameters and $\psi_{\mathrm{t-1}}$ denotes the information available at time t-1. If s_t is the sign of y_t, the joint density may be written as

$$f(y_1, \dots, y_T; \theta) = \prod_{t=1}^{T} f(y_t|s_t, \psi_{\mathrm{t-1}}; \theta) f(s_t|\psi_{\mathrm{t-1}}; \theta) \qquad (2.3.2)$$

As the observation can be obtained from its absolute value and its sign, the distribution of $|y_t|$ conditional on s_t is also valid. So the conditional density can be written as

$$\prod_{t=1}^{T} f(|y_t||s_t, \psi_{\mathrm{t-1}}; \theta) \qquad (2.3.3)$$

In this write up we apply KF in conjunction with QML to estimate the parameters of the stochastic volatility model. Using KF, the estimates of states of a model can be continuously updated on the basis of currently available information. It has also been proved that under standard conditions, the QML estimator is consistent and has a limiting normal distribution.

2.4 Testing for ARCH Effects

Let $\{\varepsilon_t\}$ be the series of residuals. The squared series $\{\varepsilon_t^2\}$ is considered to check for conditional heteroscedasticity, also known as the ARCH effects. The usual Ljung-Box statistic Q(m) is applied to the $\{\varepsilon_t^2\}$ series, where the null hypothesis is that the first m lags of autocorrelation functions of the $\{\varepsilon_t^2\}$ series are zero. The other test for conditional heteroscedasticity is the ARCH-Lagrange multiplier (ARCH-LM) test of Engle (1982). This test is equivalent to usual F-statistic for testing $H_0: a_i = 0, 1, 2, \ldots q$ in the linear regression

$$\varepsilon_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + \cdots + a_q \varepsilon_{t-q}^2 + e_t, \qquad t = q+1, \ldots, T \qquad (2.4.1)$$

where e_t denotes the error term, q is the pre-specified positive integer, and T is the sample size.

Denote

$$SSR_0 = \sum_{t=q+1}^{T} (\varepsilon_t^2 - \varpi)^2, \qquad (2.4.2)$$

where

$$\varpi = \sum_{t=q+1}^{T} \varepsilon_t^2 / T \qquad (2.4.3)$$

is the sample mean of $\{\varepsilon_t^2\}$, and

$$SSR_1 = \sum_{t=q+1}^{T} \hat{e}_t^2, \qquad (2.4.4)$$

where $\hat{e}_t$ is the least square residual.

Then, under H_0,

$$F = \frac{(SSR_0 - SSR_1)/q}{SSR_1(T-q-1)} \qquad (2.4.5)$$

is asymptotically distributed as chi-squared distribution with q degrees of freedom. The decision rule is to reject H_0 if $F > \chi_q^2(\alpha)$, where $\chi_q^2(\alpha)$ is the upper $100(1-\alpha)^{th}$ percentile of χ_q^2 or, alternatively, the p-value of F is less than α.

3. AN ILLUSTRATION

The above discussed model is applied to All-India data of monthly export of spices during the period January, 2006 to January, 2012. These are obtained from Indiastat (www.indiastat.com**)** available at I.A.S.R.I., New Delhi and the same are exhibited in Fig. 1. Out of total 73 data points, first 63 data points corresponding to the period January, 2006 to March, 2011 are used for model building and the remaining 10 data points, i.e. from April, 2011 to January, 2012 are used for validation purpose. A perusal of Fig. 1, appended as an annexure 1, indicates presence of volatility at several time-epochs.

A high volatility is noticed in March, 2007 where export suddenly jumped almost 140% to the level of Rs. 402 crores and then an abrupt dip in the very next month to Rs. 301 crores. Similar type of jump is noticed at time-epochs, like March, 2008. Volatility can also be seen in many time-epochs like August, 2007, October, 2009, March, 2010, and December, 2010.

Firstly, the appropriate ARIMA model is chosen on the basis of minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values given as

$$AIC = Tlog(\sigma^2) + 2(p + q + 1) \qquad (3.1)$$

$$BIC = Tlog(\sigma^2) + (p + q + 1)logT \qquad (3.2)$$

On the basis of aforementioned criteria, the ARIMA(*1*,*1*,*0*) model is selected for modelling of the monthly export of spices.

Table 1. Estimates of parameters along with their standard errors for fitted ARIMA model

Parameter	Estimate	Standard error
Intercept	11.56	6.51
AR1	-0.25	0.12

The Acf of the squared residuals of the fitted ARIMA(*1*,*1*,*0*) model is found to be reasonably high at lag 6, which is *-0.22*. Consequently, the ARCH-LM test is carried out at lag 6 to check for conditional heteroscedasticity. The ARCH-LM test statistic at lag 6 is computed using eq. (2.4.5) and found to be significant at 5% level. But it is not reasonable to apply ARCH model of order 6 in view of the enormously large number of parameters. To this end, a parsimonious model is needed in predicting the conditional variances. Consequently, on the basis of minimum AIC and BIC values the AR(*1*)-GARCH(*1*,*1*) model is selected for modelling the data under consideration. The AIC and BIC values for GARCH model with Gaussian distributed errors can be calculated by

$$AIC = \sum_{t=v}^{T}(logh_t + \varepsilon_t^2 \mathrm{h}_t^{-1}) + 2(p + q + 1) \qquad (3.3)$$

and

$$BIC = \sum_{t=v}^{T}(logh_t + \varepsilon_t^2 \mathrm{h}_t^{-1}) + 2(p + q + 1)\log{(T - v + 1)} \qquad (3.4)$$

In the present investigation, the Gaussian maximum likelihood estimation procedure available in EViews software package, Ver. 4 is used for data analysis. The fitted model is given by

$$y_t = \underset{(145.01)}{593.75} + \underset{(0.06)}{0.91}y_{t-1} + \varepsilon_t$$

where $\varepsilon_t = h_t^{1/2}\xi_t$, and h_t satisfies the variance equation

$$h_t = \underset{(5345.63)}{14683.53} + \underset{(0.24)}{0.31}\varepsilon_{t-1}^2 - \underset{(0.06)}{1.02}h_{t-1}$$

However, the GARCH assumption that the volatility is driven by past observable variables only can become a constraint. So, the SV model is fitted to the data under consideration. Using the one-step ahead prediction error and its corresponding mean-squared error obtained via KF, the prediction error decomposition form of the likelihood is obtained. Subsequently, the unknown parameters θ is estimated by maximizing the likelihood. Applying the steps mentioned, the fitted SV model is given as

$$\log(y_t^2) = 4.79 + h_t + \xi_t$$

$$h_{t+1} = (0.96 - 0.54\, s_t/0.53)h_t + s_t\{0.55 + 0.54/0.53\,(logy_t^2 - 2.33)\} + \eta_t^+$$

The parameters σ_η^2 and φ obtained by maximizing the prediction error decomposition form of the likelihood are quite close to the σ_η^2 obtained by posterior analysis of innovation based on h_{t+1} and its lag value, when the information till time *t* is known. This validates the estimates of the parameters.

To study the appropriateness of the fitted SV model, the autocorrelation function of the standardized residuals and squared standardized residuals are computed and reported in tables 3 and 4 respectively. It is found that, in both situations, the autocorrelation function is not significant at 5% level, thereby confirming that the mean and variance equations are correctly specified. The AIC and BIC values computed for fitted SV model are respectively *522.20* and *526.48*, which are much lower than the corresponding values, viz. 536.30 and *547.01* for the fitted AR(1)-GARCH(*1*,*1*) model.

Also, the performance of fitted SV and GARCH models is evaluated using the Mean Square Error (MSE) criterion, defined as

$$\mathrm{MSE} = 1/\mathrm{N}\ \sum_{\mathrm{T}=1}^{\mathrm{N}}\{\mathrm{Y_T} - \widehat{\mathrm{Y}}_\mathrm{T}\}^2 \qquad (3.5)$$

The MSE values for fitted SV and GARCH models are respectively computed as 3616.36 and 3905.27. Thus, all this indicates that the fitted SV model has performed better than the AR(1)-GARCH(1,1) model for modelling the volatile data under consideration. The graph of fitted SV model along with data points is exhibited in Fig. 2, which indicates that the fitted SV model is able to capture the volatilities present in the data to a reasonable extend. Conditional standard deviation for the fitted model is plotted in Fig. 3.

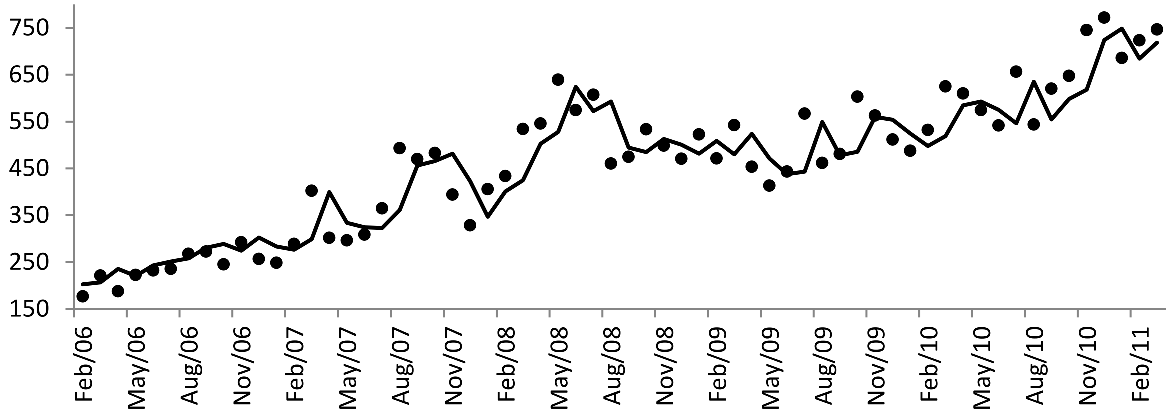


Fig. 2. Fitted Stochastic Volatility Model along with data points

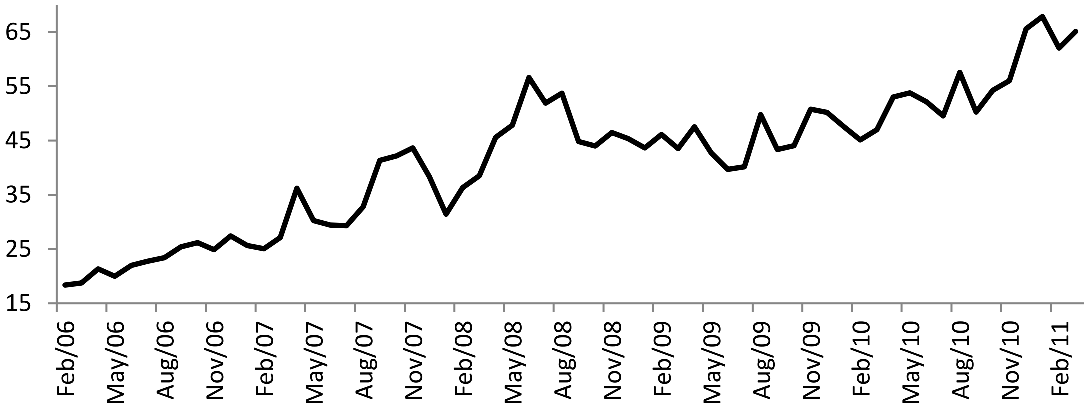


Fig 3. Conditional standard deviation of fitted Stochastic volatility model

3.2 Forecasting performance

In this sub-section, GARCH and SV model is compared on their ability to forecast. We take 10 data points corresponding to All-India data of monthly export of Spices from April, 2011 to January, 2012 as hold-out-data. One-step ahead forecasts are computed along with the corresponding forecast standard errors and reported in Table 2.

Table 2. One-step ahead forecasts of export data (in Rs. Crore)

		SV model	AR(1) - GARCH(1,1) model
Months	Actual	Forecast	Forecast
Apr-11	758.45	801.72 (65.94)	733.03 (72.47)
May-11	890.10	872.83 (67.21)	743.83 (118.21)
Jun-11	876.86	903.91 (74.88)	863.82 (99.55)
Jul-11	1007.94	998.92 (77.00)	851.74 (96.89)
Aug-11	1222.66	1040.95 (84.93)	971.20 (121.62)
Sep-11	1248.52	1225.94 (99.33)	1166.89 (137.33)
Oct-11	1266.68	1297.28 (106.30)	1190.44 (23.03)
Nov-11	1160.27	1229.62 (109.83)	1207.00 (137.53)
Dec-11	1256.98	1179.96 (105.07)	1110.03 (58.56)
Jan-12	1071.73	1158.72 (108.84)	1198.15 (137.76)

An inspection of the table indicates that SV model performs comparatively well. The performance of fitted models is also compared on the basis of one-step-ahead Mean square prediction error (MSPE), Mean absolute prediction error (MAPE) and Relative mean absolute prediction error (RMAPE) given as

$$\mathrm{MSPE} = 1/\mathrm{N}\ \sum_{\mathrm{i}=0}^{\mathrm{N}-1}\left\{\mathrm{Y}_{\mathrm{T+i+1}} - \widehat{\mathrm{Y}}_{\mathrm{T+i+1}}\right\}^2 \qquad (3.2.1)$$

$$\mathrm{MAPE} = {}^{1}/_{\mathrm{N}} \sum_{\mathrm{i}=0}^{\mathrm{N}-1}\left\{\left|\mathrm{Y}_{\mathrm{T+i+1}} - \widehat{\mathrm{Y}}_{\mathrm{T+i+1}}\right|\right\} \qquad (3.2.2)$$

$$\mathrm{RMAPE} = {}^{1}/_{\mathrm{N}} \sum_{\mathrm{i}=0}^{\mathrm{N}-1}\left\{ {\left|\mathrm{Y}_{\mathrm{T+i+1}} - \widehat{\mathrm{Y}}_{\mathrm{T+i+1}}\right|}\Big/{\mathrm{Y}_{\mathrm{T+i+1}}} \right\} \times 100 \qquad (3.2.3)$$

The MSPE, MAPE and RMAPE values for fitted SV model are respectively computed as 5575.31, 56.47 and 5.09, which are found to be lower than the corresponding ones for fitted AR(1)-GARCH(1,1) model, viz. 16206.8, 107.02 and 9.73 respectively. This indicates the superiority of SV model over GARCH model for forecasting purposes.

To sum up, the SV model has performed satisfactorily for modelling as well as forecasting of the volatile data under consideration.

4. CONCLUSION

In this write up, Stochastic volatility (SV) model and its properties is thoroughly studied. The methodology for estimation of SV model using Kalman filter in conjunction with Quasi-maximum likelihood method is also discussed. As an illustration, modelling and forecasting of volatile All-India spices monthly export data is carried out. Superiority of SV model over GARCH model for the data under consideration is clearly demonstrated for modelling and forecasting. As future work, possibility of application of Monte Carlo technique using Particle filtering approach may be explored for parameter estimation and volatility forecasting.

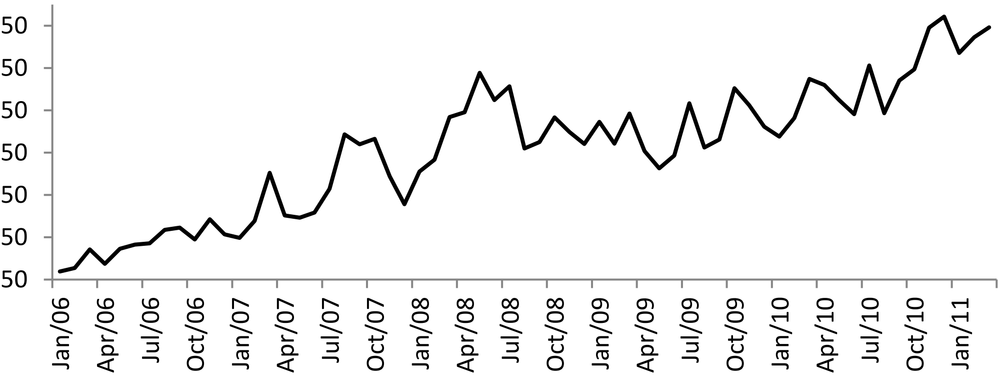


Fig. 1. Plots of All-India data of monthly export of spices

Table 3. Autocorrelation functions and partial autocorrelation functions of the standardized residuals for the fitted Stochastic volatility model

	ACF	PACF	Q-Stat.	Prob.
1	-0.147	-0.147	1.3841	0.239
2	-0.100	-0.125	2.0397	0.361
3	0.166	0.136	3.8729	0.276
4	0.000	0.037	3.8729	0.423
5	0.220	0.271	7.1899	0.207
6	-0.206	-0.172	10.150	0.118
7	-0.097	-0.123	10.819	0.147
8	0.134	-0.022	12.120	0.146
9	0.058	0.129	12.370	0.193
10	-0.148	-0.131	14.029	0.172
11	-0.014	0.046	14.045	0.231
12	0.199	0.176	17.150	0.144
13	-0.079	-0.067	17.654	0.171
14	0.037	0.027	17.764	0.218
15	-0.133	-0.125	19.253	0.202
16	0.064	0.027	19.603	0.239
17	0.062	-0.096	19.935	0.278
18	-0.163	0.010	22.318	0.218
19	0.033	0.013	22.416	0.264
20	-0.084	-0.093	23.084	0.285

Table 4. Autocorrelation functions and partial autocorrelation functions of the squared standardized residuals for the fitted Stochastic volatility model

	ACF	PACF	Q-Stat.	Prob.
1	-0.047	-0.047	0.1444	0.704
2	-0.080	-0.082	0.5612	0.755
3	-0.146	-0.155	1.9645	0.580
4	-0.056	-0.083	2.1758	0.703
5	-0.022	-0.060	2.2092	0.820
6	-0.054	-0.100	2.4123	0.878
7	0.040	-0.002	2.5242	0.925
8	0.214	0.194	5.8398	0.665
9	0.192	0.221	8.5778	0.477
10	-0.096	-0.019	9.2664	0.507
11	-0.264	-0.201	14.617	0.201
12	-0.051	-0.037	14.825	0.251
13	-0.047	-0.083	15.002	0.307
14	-0.100	-0.194	15.823	0.324
15	-0.072	-0.188	16.251	0.366
16	0.207	0.086	19.922	0.224
17	-0.078	-0.239	20.455	0.252
18	-0.042	-0.153	20.610	0.300
19	-0.105	-0.010	21.619	0.304
20	-0.050	0.007	21.855	0.348

An Introduction to Genetic Algorithm

Kanchan Sinha, K.N. Singh, Mrinmoy Ray and Achal Lama
ICAR-IASRI, New Delhi

Introduction

Genetic Algorithm (GA) is a search-based optimization technique based on the principles of Genetics and Natural Selection. The algorithm performs a search in providing an optimal solution for evaluation (fitness) function of an optimization problem. GAs deal simultaneously with multiple solutions and use only the fitness function values. John Holland introduced genetic algorithm in 1960 based on the concept of Darwin's theory of evolution; afterwards, his student David E. Goldberg extended GA in 1989. The process of natural selection starts with the selection of fittest individuals from a population. They produce offspring which inherit the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance at surviving. This process keeps on iterating and at the end, a generation with the fittest individuals will be found.

1. Basic Terminology

GA works in an iterative manner by generating new populations of strings from old ones. Every string is the encoded binary, real, etc. which is known as chromosome. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. To understand GAs, it is required to keep in mind the basic terminology.

Individual- Any possible solution.

Population- Group of all individuals.

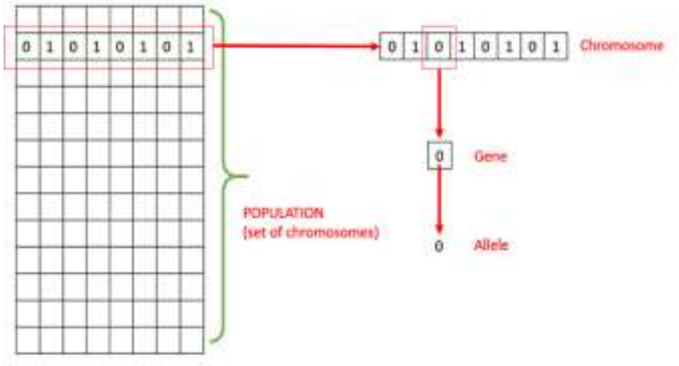
Chromosomes- A chromosome is one such solution for a given problem.

Gene- A gene is one element position to the given problem.

Allele- It is the value a gene takes for a particular chromosome.

Locus- The position of a gene on the chromosome.

Genome- Collection of all chromosomes for an individual.



Genotype- The set of genes representing the chromosome.

Phenotype-The actual physical representation of the chromosome.

Decoding and Encoding- For simple problems, the phenotype and genotype spaces are the same. However, in most of the cases, the phenotype and genotype spaces are different. Decoding is a process of transforming a solution from the genotype to the phenotype space, while encoding is a process of transforming from the phenotype to the genotype space. Decoding should be fast as it is carried out repeatedly in a GA during the fitness value calculation.

2. Basic Principles of a GA-based optimization technique

Once the problem is encoded in a chromosomal manner and a fitness measure for discriminating good solutions from bad ones has been chosen, the process starts to evolve solutions to the search problem using the following steps

i. Initialization: Initial population is randomly selected. Population size is the number of chromosomes in each generation and it is an important parameter to increase the performance of genetic algorithms. There is no standard to specify the size.

ii. Fitness Function: The fitness function determines how fit an individual is (the ability of an individual to compare with other individuals). After representing each chromosome, the right way to serve to search space, next is to calculate the fitness value of each individual. The process of calculating the fitness value of a chromosome is called evaluation.

iii. Selection: Selection is a significant part of the evolutionary algorithm to reach the best chromosomes. The selection operator chooses chromosomes from mating pool according to GAs working principle, “the fittest individuals have a greater chance of survival than weaker ones”.

iv. Crossover: Crossover operator provides new offspring for the next generation with exchanging information between randomly selected two parent chromosomes. Diversification is very important in GA and crossover provides much superiorities to GA in terms of exploration and diversification abilities to achieve global optimum point.

v. Mutation: Mutation operator is utilized to put new genetic information with modifying genes of a chromosome selected with a mutation probability. Mutation is a divergence operation which provides avoiding local optima in the search space.

vi. Stopping criteria: The stopping criteria is decided according to the improvement of the fitness function. The individual with high fitness remain and the ones with low fitness are removed. If there is no improvement on the last improved solution’s fitness function after a prescribed number of iteration, then the algorithm is stopped.

The genetic algorithm procedure is depicted in the following figure.

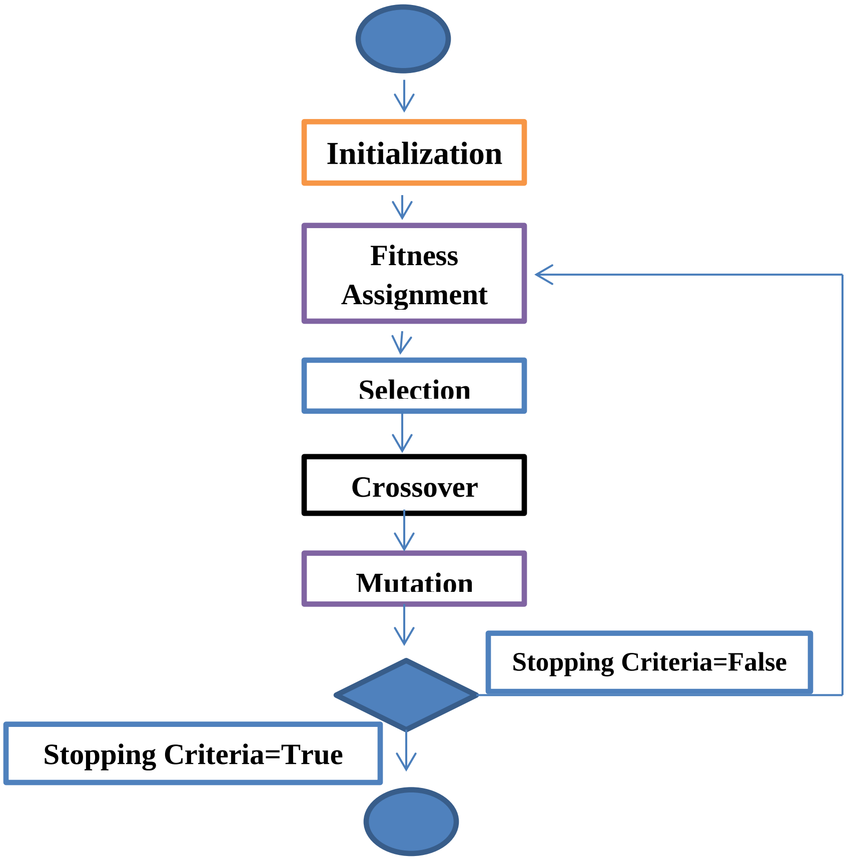


Fig.: Genetic Algorithm

3. Advantages of GAs

- Does not require any derivative information (which may not be available for many real life problems)
- Is faster and more efficient as compared to the traditional methods.
- Has very good parallel capabilities.
- Optimize both continuous and discrete functions and multi-objective problems.
- Provides a list of good solutions and not just a single solution.
- Useful when search space is very large and there are a large number of parameters involved.

4. Limitations of GA

Like other techniques, GAs also suffer from a few limitations. These include

- GAs are not suited for all problems, especially problems which are simple and for which derivative information is available.
- Fitness value is calculated repeatedly which might be computationally expensive for some problems.
- Being Stochastic, there are no guarantee on the optimality or the quality of the solution.
- If not implemented properly, the GA may not converge to the optimal solution.

5. Application of GAs

Genetic Algorithms are primarily used in optimization problems of various kinds, but they are frequently used in various domain as well. The different applications of GAs are listed below.

- Natural Sciences, Mathematics and Computer Science
- Earth Sciences
- Finance and Economics
- Social Sciences
- Industry, Management and Engineering
- Biological Sciences and Bioinformatics, etc.

6. Suggested Readings

Goldberg, David (1989) Genetic algorithms in search, optimization and machine learning, Reading, MA: Addison-Wesley Professional. ISBN 978-0201157673.

Holland, J. H. (1992) Adaption in natural and artificial systems, MIT press, Cambridge.

Melanie, Mitchell (1999) An introduction to genetic algorithm, fifth printing, MIT press, Cambridge.

Randy, L. H. and Haupt, S. E. (2004) Practical genetic algorithms, Wiley, Second Edition.

ARFIMA Models for Modeling and Forecasting Long Memory Time Series

Achal Lama, Santosha Rathod and R S Shekhawat
ICAR-IASRI, New Delhi
achal.lama@icar.gov.in

Introduction:

Time series forecasting is an important area of forecasting in which successive measurements are made over a period of time for the same variable and analyzed to develop a model describing the underlying relationship. This approach is particularly useful when there is little or no satisfactory knowledge about the explanatory or prediction variable is available. One of the most important and widely used time series model is the autoregressive integrated moving average (ARIMA) model. The popularity of the Auto Regressive Integrated Moving Average (ARIMA) model is due to its statistical properties as well as the well-known Box–Jenkins methodology (Box and Jenkins 1970) in the model building process.

The autocorrelation of the time series is expected to be decrease or vanish as the observations are distance apart in time for example in ARIMA model the autocorrelation decreases exponentially as the time lag increases and in some series the decay can occur at much slower hyperbolic rate. Such series are said to have long memory and are commonly prevail in stock market prices and in economic time series such as stock price, economic growth rate, inflation rate, oil price, agricultural commodity price and GDP figures *etc.*, the time series showed a characteristic of "long memory faculty' (Xu 2010 and Paul 2014). Long memory shows that time series exists a continuous long-term dependency among the distant time interval measurement. When the delay order number k was larger, time series has a correlation in the time value and t-k time value, and this is often measured by the autocorrelation coefficient of the series, and the memory extend of the series can be judged by reduction in the autocorrelation coefficient curve.

A popular class of models for time series with long memory behavior is the ARFIMA model. This kind of models extended classical ARIMA models by assuming the differencing parameter *d* as a real value. It is well known that ARFIMA models are linear time series model.

Since some long memory time series have both linear and non-linear structures, ARFIMA models can be inadequate for this type of series.

Long memory process:

Natural phenomenon and social economic phenomenon are both regular dialectical development courses. All movement has certain inertia, and it shows a kind of dynamic of the system, namely memory (Hurst 1951). After Hurst found long memory of hydrological time series from the tidal data, the research of long memory has been caused widely public concern such as fluid science meteorology and geophysics and so on. Long memory in time-series can be defined as autocorrelation at long lags (Robinson 1995).

Definition: we assume that the time series X_t has autocorrelation function ρ_t and *t* is the lag number. If ρ_t satisfies the condition:

$$\lim_{T\to\infty} \sum_{t=-n}^{n} |\rho_t| \to \infty \tag{1}$$

Then is X_t called long memory time series.

For checking the presence of long memory of the data the statistical methods commonly used are correlation coefficient method, the classical R/S test (or heavy rescaled range test), the modified R/S test (MR/S), KPSS method, logarithmic diagram method (GPH), and Gauss semi-parametric estimation method (GSP) (Aarthi 2012).

R/S analysis method:

Let us consider the time series X_t of the sample length *T* is divided into *k* son intervals of length *n*(*n´k* = *T*) , and the average of *n* series observed values is $\bar{X}_{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} x_i$ The range of each subinterval is defined *R*(*n*) , and the standard deviation *S*(*n*).

$$Q_n = \frac{R(n)}{S(n)} \tag{2}$$

The range *R(n)* and Standard Deviation *S(n)* are respectively,

$$R(n) = \max_{1\le k\le n} \sum_{j=1}^{k} (x_j - \overline{x_n}) \min_{1\le k\le n} \sum_{j=1}^{k} (x_j - \overline{x_n}) \tag{3}$$

$$S(n) = \left|\frac{1}{n} \max_{1\le k\le n} \sum_{j=1}^{k} (x_j - \overline{x_n})^2\right|^{1/2} \tag{4}$$

We can prove that $\lim_{n\to\infty} n^{-H}\, Q_n = C$ is a constant, and H is Hurst index, so we can get approximate estimate of $H = \ln Q_n \,/\, \ln n$. In general, the R/S analysis method is described as follows.

$$(R/S)_n = C.n^H \tag{5}$$

In eqn. (5), *R* is rescaled range, *S* is the standard deviation, *H* is Hurst index, *C* is a constant, and *n* is sample observation number. On the (5) eqn. of logarithmic, we get

$$log(R/S)_n = \log(C) + \ H \log{(n)} \tag{6}$$

Hurst index is in [0, 1] value, according to its value the time series can be divided into three different types:

1. H=0.5 indicating the correlation coefficient between past and future increment of series is zero, and namely it does not affect the future, and the incremental process is an independent random process, the series is the random walk and it is gradual process.
2. 0<H<0.5 indicating a reverse persistent series (mean reversion) then the autocorrelation coefficient is between 0.5 and 1 and When *H* is closer to 0 this process has more frequent reversible.
3. 0.5<*H*<1 indicating persistent or trend enhanced series. Then the autocorrelation coefficient is between 0 and 1. This indicates that if the past has the trend of growth, it means that this trend will continue in the future. *H* is closer to 1, and the trend is more obvious; *H* is closer to 0.5, the trend is gradually becoming random.

The ARFIMA model:

ARFIMA models are used to model long range dependent time series. ARFIMA models were introduced by Granger and Joyeux (Granger and Joyeux 1980). ARFIMA model is expressed as follows;

$$\varphi(B)(1-B)^d X_t = \theta(B)e_t \text{ , -0.5<}d\text{<0.5} \tag{7}$$

Where, *B* is the back-shift operator such that $BX_t=X_{t-1}$ and e_t is a white noise process with *E(ei)*=0 and variance $\sigma_e{}^2$. The polynomials $\varphi(B) = (1 - \varphi_1\, B - \cdots .. - \varphi_p B^p$and $\theta(B) = (1 - \theta_1\, B - \cdots .. - \theta_q B^q$ have orders *p* and *q* respectively with all their roots outside the unit circle., Beran extended the estimation of ARFIMA models by considering the following variation of the ARFIMA model (Beran 1995)

$\varphi(B)(1-B)^d(1-B)^m X_t = \theta(B)e_t$, -0.5<d<0.5 (8)

The integer m is the number of times that must be differenced to achieve stationary, and thus differencing parameter is given by d=d+m.

Studies about the parameter estimation of ARFIMA models still continue. Many maximum likelihood (ML) methods for ARFIMA are proposed in literature such as approximate ML methods (AML) by Beran; exact ML method (EML) (Sowell 1992), conditional sum of square (CSS) method by Chung and Baillie (Chung and Baillie 1993). Note that CSS method is as efficient as EML method and it is identical with AML method by Beran, that is based on infinity autoregressive presentation.

Beran, has given some properties of a long memory stationary series as follows (Beran 1995):

1. Certain persistence exists. The observations tend to stay at high levels in some periods, and at low levels in some other periods.
2. During short-time periods, there seems to be periodic cycles. However, looking through the whole process, no apparent periodic cycles could be identified.
3. Overall, the process looks stationary.

Quantitatively, for a stationary process, these features could be described as

1. When adding more observations, the variance of the sample mean and variance
2. decays to zero at a slower rate than n^{-1} which is the rate at which a white noise decays, and is asymptotically equal to a constant g times n^{-c} for some *0* < c < *1*.
3. The correlation $r_j$decays to zero slowly and is asymptotically equal to a constant time $j^{ca}$for some *0* < c < *1*.

The ARFIMA model building has same procedure of model building as Box-Jenkins ARIMA methodology.

Illustration (Rathod *et al*, 2016):

For the present study, the daily spot price (Rupees/Quintal) of agricultural commodity; mustard in Mumbai market for the period 1st January, 2009 to 14th February, 2012 are used. The data is collected form Ministry of Consumer's Affairs, Government of India. Out of 1140 total observations, 1080 have been used for model estimation and remaining 60 observations are used for validation. Summary statistics of mustard spot price is given in Table 1 and time series plot of above dataset has been exhibited in Fig. 3. Graphically, the plot indicates that the dataset is stationary. Further to validate the stationarity of the series, two tests namely Augmented Dickey-Fuller test and Philips-Peron test are used. Results of the stationarity tests are reported in table 2. The result indicates that spot price time series data of mustard in Mumbai is stationary.

Table 1: Summary statistics of mustard spot price

Statistic	Series	Statistic	Series
Observation	1140	Standard Deviation	320.41
Mean	2849.89	Kurtosis	1.60
Median	2900.00	Skewness	-0.75
Mode	2750.00	Coefficient of Variation (%)	11.24

Table 2: Testing for stationarity

ADF test statistic				**PP test statistic**			
Single mean	With trend	Probability		Single mean	With trend	Probability	
		Single mean	With trend			Single mean	With trend
-6.24	-8.44	<0.001	<0.001	-7.65	-7.81	<0.001	<0.001

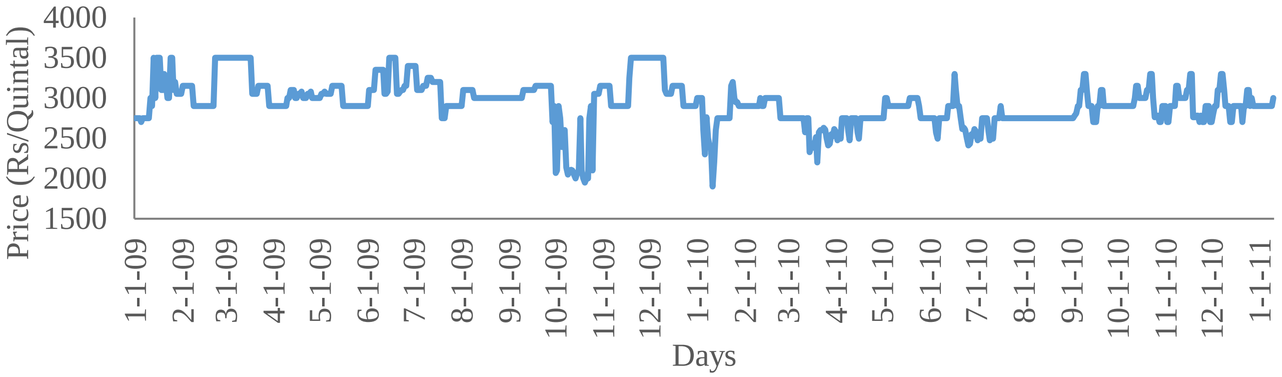


Fig. 3: Time series plot of actual series.

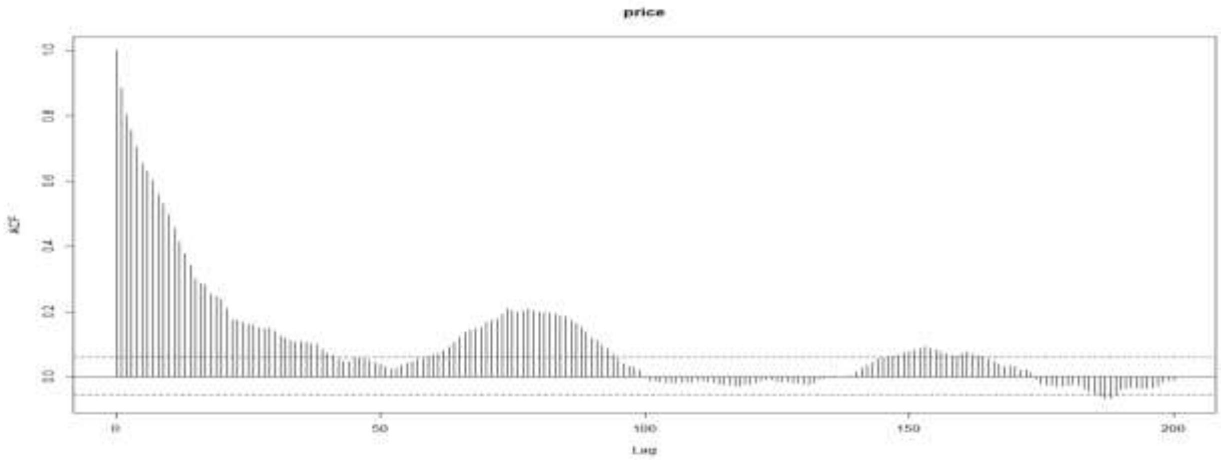


Fig. 4: Plot of ACF of actual series.

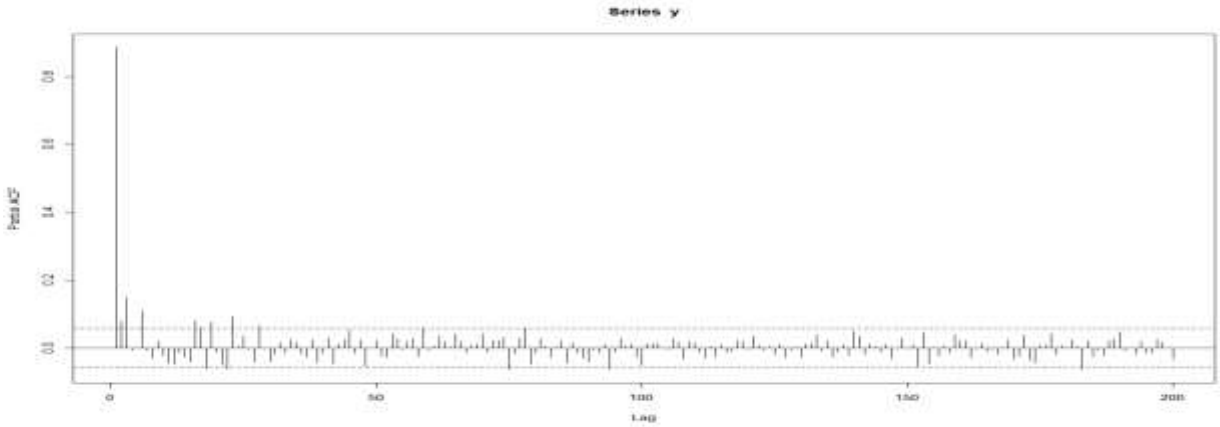


Fig. 5: Plot of PACF of actual series.

Figures 4 and 5 depicts the plot of autocorrelation function (ACF) and partial autocorrelation function (PACF) for the actual price series. Though the stationarity tests validated that the series is stationary, but plot of ACF shows a slow decay towards zero indicating the possible presence of long memory. Therefore, presence of long memory is tested and parameter *d* has been estimated using GPH method and it was obtained as 0.40 (0.13) have been used to estimate the long memory parameter.

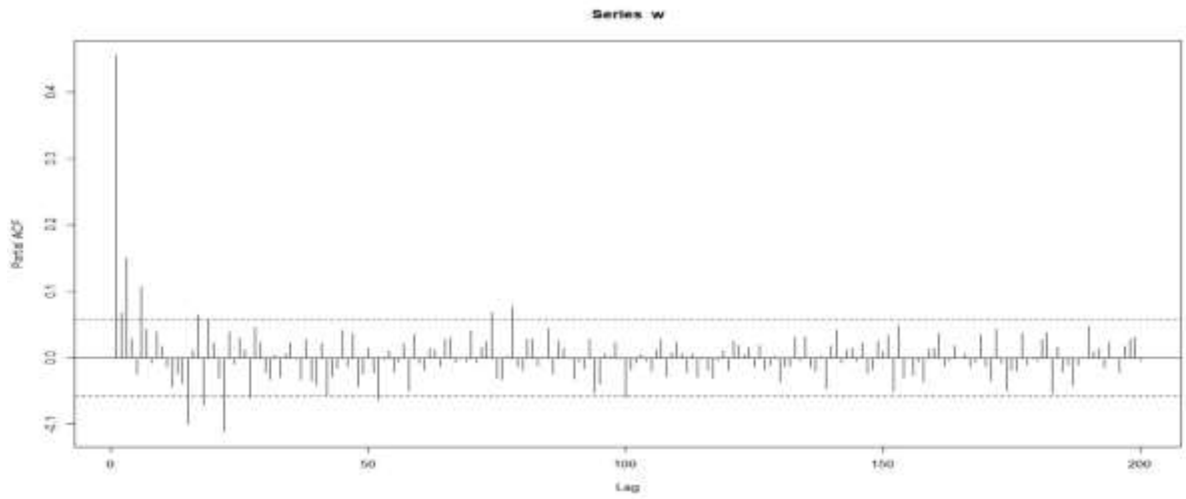


Fig. 6: Plot of ACF of fractional differenced series

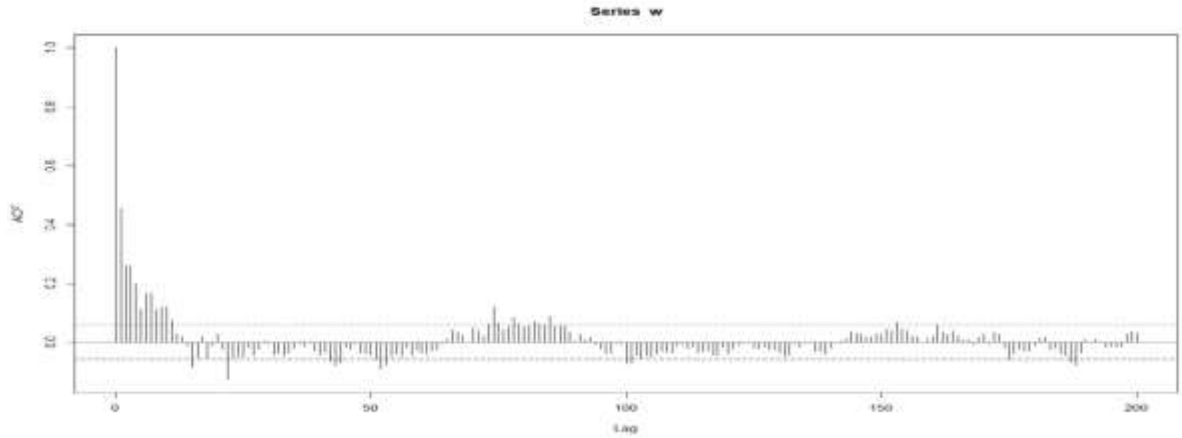


Fig. 7: Plot of ACF of fractional differenced series

The fractional differenced series with parameter (*d*) as 0.40 is computed. Based on graphical representation of ACF and PACF of fractionally differenced series, one can say, the decay rate of ACF has improved as compared to the decay of ACF in actual series (Fig. 6 and 7). Here we estimated different ARFIMA specifications for the data under consideration.

Table 4: ARFIMA model parameters

Parameters	**Estimates**	**Std. Error**	***z*-value**	**Pr(>|z|)**
AR1	0.980	0.008	115.86	<0.001
MA1	-0.128	0.055	-2.30	0.020

The MAPE (Mean Absolute Percentage Error) obtained for training and testing data set is 3.58 and 5.41). Long memory time series has been analyzed by using ARFIMA models which are based on linear structure. However, it is not absolutely certain that the over difference is the only reason which degrade the performance of the ARIMA model in case of long memory time series data

Bibliography:

Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short and long memory ARIMA models, *Journal of Royal Statistical Society, Series B*, 57(4): 659-672.

Box, G.E.P. and Jenkins, G. (1970). Time series analysis, Forecasting and control, Holden-Day, San Francisco, CA.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers*, 35(16): 700-779.

Paul, R.K. (2014). Forecasting Wholesale Price of Pigeon Pea Using Long Memory Time-Series Models, *Agricultural Economics Research Review*, 27 (2):167-176.

Aarthi, R, S., Muralidharan, D, Swaminathan, P. (2012). Double Compression of Test Data Using Huffman Code, *Journal of Theoretical and Applied Information Technology*, 39(2): 104-113.

Rathod, S., Singh, K, N., Paul, R.K., Meher, R.K., Mishra, G.C., Gurung, B., Ray, M. and Sinha, K. 2017. An Improved ARFIMA Model using Maximum Overlap Discrete Wavelet Transform (MODWT) and ANN for Forecasting Agricultural Commodity Price. *Journal of the Indian Society of agricultural Statistics*. **71(2)**: 103–111.

Robinson, P.M. (1995). Log-periodogram regression of time series with long-range dependence. *The Annals of Statistics,* 23: 1048-1072.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *Journal of Econometrics,* 53:165-188.

Xu, Y.C. (2010). The effectiveness of long memory of Financial Time series R/S test, *China Management Science,* 11(18): 204-208.

An Introduction to Recurrent Neural Networks

Kanchan Sinha
ICAR-IASRI, New Delhi

1. Introduction

Artificial neural networks (ANNs) are computational methods that mimic the behaviour of the human brain's central nervous system that are made from layers of connected units called artificial neurons. Neural network is a class of generalized non-linear, nonparametric, data driven approach can be viewed as a powerful learning models that achieve state-of-the-art results in a wide range of supervised and unsupervised machine learning tasks. A general neural networks architecture consists of an input layer that accepts external information, one or more hidden or middle layer that provide non-linearity to the model and an output layer that provides the target value. Each layer contains one or more nodes. All the layers in a multi-layer neural network are connected through an acyclic arc. A neural network model with p number of input nodes and q number of hidden nodes consists of $q(p+2)+1$ number of parameters. As the number of layers increases, the complexity of networks (number of parameters) increases too. More number of layers or recurrent connections generally increases the depth of the network and empowers it to provide various levels of data representation and feature extraction referred to as "deep learning". An artificial neural network with recurrent connection is called as recurrent neural networks (RNNs) which are capable of learning features, modelling sequential data for sequence recognition and prediction. Recurrent neural networks (RNNs) are a kind of neural network that specialize in processing sequences. They are often used in Natural Language Processing (NLP) tasks because of their effectiveness in handling text. Recurrent neural networks are made of high dimensional hidden states with non-linear dynamics. The structure of hidden states works as the memory of the network and state of the hidden layer at a time is conditioned on its previous state. The type of structure enables the RNNs to store, remember, and process past complex signals for long time periods. RNNs can map an input sequence to the output sequence at the current time step and predict the sequence in the next time step.

2. Artificial Neural Network

Single hidden layer feed forward network is the most popular for time series modelling and forecasting. This model is characterized by a network of three layers of simple processing units, and thus termed as multilayer ANNs. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer.

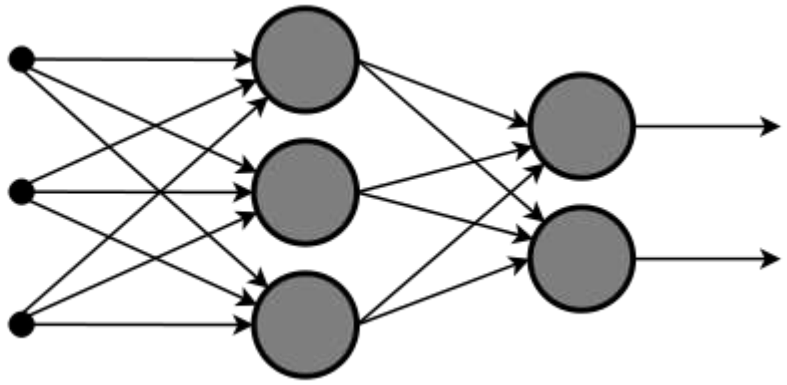


Fig. 1: Three layers feed forward networks

In this network the information moves in only one direction, forward, from the input nodes, through the hidden nodes and to the output nodes. There are no cycles or loops in the network. The relationship between the output (y_t) and the inputs ($y_{t-1}, y_{t-2}, ..., y_{t-p}$) can be mathematically represented as follows:

$$y_t = f\left\{\sum_{j=0}^{q} \omega_j g\left(\sum_{i=0}^{p} \omega_{ij} y_{t-i}\right)\right\} \qquad \text{(i)}$$

where, ω_j (*j* = 0,1,2,…. . , *q*) and ω_{ij} (*i*= 0,1,2,…, *p*, *j*= 0,1,2,...*q*) are the model parameters often called the connection weights, *p* is the number of input nodes and *q* is the number of hidden nodes, *g* and *f* denote the activation function at hidden and output layer respectively. Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity. Training of a neural network involves the following steps:

i. Input a dataset.

ii. The network will take the dataset and apply some complex computations to it using randomly initialized variables (called weights and biases).

iii. A predicted result will be produced.

iv. An error will be obtained after comparing the result to the actual value.

v. Propagating the error back through the same path will adjust the variables.

vi. Steps i-v are repeated until it is confident to say that the variables are well defined.

vii. A prediction is made by applying these variables to a new unseen input.

3. Recurrent neural networks

Recurrent neural networks are networks with loops in them, allowing information to persist. It is able to 'memorize' parts of the inputs and use them to make accurate predictions. These networks are the heart of speech recognition, language modelling, translation, image captioning and more. A chunk of neural network, looks at some input and outputs a value. A loop allows information to be passed from one step of the network to the next.

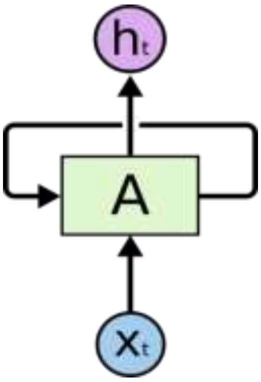


Fig. 2: Recurrent Neural Networks have loops

In traditional neural network, inputs and outputs are considered as independent of each other. As the sequential pattern exist in time series data, such a neural network does not give efficient results for the time series forecasting. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor. They're the natural architecture of neural network to use for such data. As an alternative network, RNN is more efficient to learn the dependency between observations. The simple architecture and unrolled version of RNNs is shown in figure 3. The simple RNN is a network with loops which allows persisting information to be passed from one step of the network to the next. In the diagram for the time steps $0,1,2,\ldots,t;\ x_0,x_1,x_2,\ldots,x_t$ are the inputs, A is the hidden state and $h_0,h_1,h_2,\ldots,h_t$ are the outputs. A_t hidden state is an activation function (normally *tanh*) which takes its input from the hidden state of the previous step A_{t-1} and the output of the current step x_t.

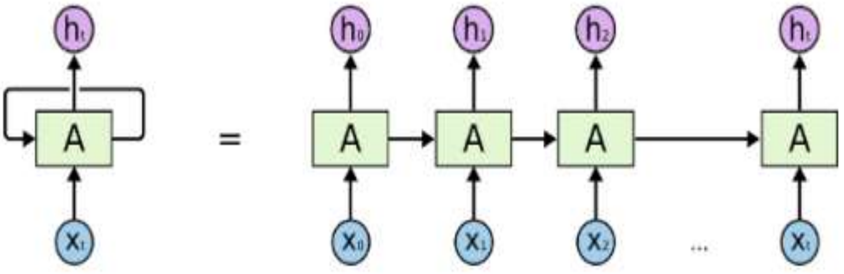


Fig. 3: An unrolled recurrent neural networks

This process is described by the following equation

$A_t = f(A_{t-1}, x_t)$ (ii)

RNNs use backpropagation through time (BPTT) to optimize weights during training by using the chain rule to go back from the latest time step to the previous steps. Figure 4 presents a typical RNN looks like which is being unfolded or unrolled into a full network. By unrolling, it means to write out the network for the complete sequence. For example, if the sequence is a sentence of five words, the network would be unrolled into five-layer neural network, one layer for each word. The formulas that govern the computation happening in a RNN are as follows:

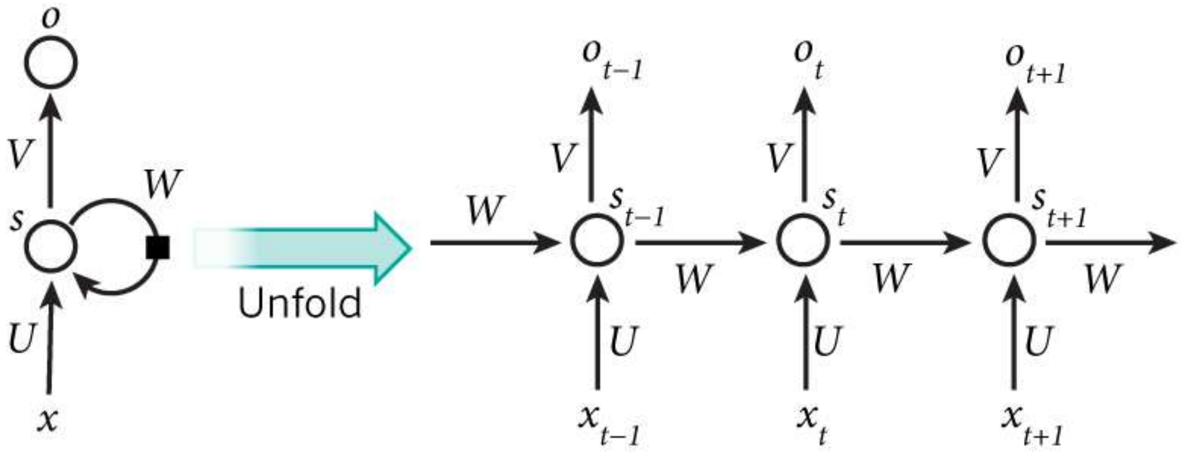


Fig. 4: A recurrent neural network and the unfolding in time of the computation involves in its forward computation

x_t is the input at time step t. s_t is the hidden state at time step t. It is the “memory” of the network. s_t is calculated based on the previous hidden state and the input at the current step as

$s_t = f(Ux_t + Ws_{t-1})$ (iii)

The function f is usually nonlinear such as $tanh$ or $ReLU$. s_{-1} which is required to calculate the first hidden state, is typically initialized to all zeroes. o_t is the output at time step t. For example, to predict the next word in a sentence it would be a vector of probabilities across our vocabulary.

$$o_t = softmax(Vs_t) \tag{iv}$$

3.1 Salient points to remember for RNNs

- Hidden state s_t as the memory of the network, captures information about what happened in all the previous time steps. The output o_t is calculated solely based on the memory at time.

- Unlike a traditional deep neural network, which uses different parameters at each layer, a RNN shares the same parameters (U, V, W) across all steps as mentioned in the figure 4. This reflects the performing of same task at each steps just with different inputs which reduces the total number of parameters need to learn.

- In the figure 4, there is output at each time step which may not be necessary depending on the task. Similarly inputs at each time step is not necessary. The main features of RNNs is its hidden state, which captures some information about a sequence.

3.2 Application of RNNs

3.2.1 Language modelling and prediction

In this method, the likelihood of a word in a sentence is considered. The probability of the output of a particular time-step is used to sample the words in the next iteration (memory). In language modelling, input is usually a sequence of words from the data and output will be a sequence of predicted word by the model. While training the output of the previous time step will be the input of the present time step.

3.2.2 Speech Recognition

A set of inputs containing phoneme (acoustic signals) from an audio is used as an input. This network will compute the phonemes and produce a phonetic segments with the likelihood of output.

3.2.3 Machine Translation

In machine translation, the input will be the source language (e.g. Hindi) and the output will be in the target language (e.g. English). The main difference between machine translation and language modelling is that the output starts only after the complete input has been fed into the network.

3.2.4 Image Recognition and Characterization

Recurrent neural network along with a Convolutional Neural Network work together to recognize an image and give a description about it if is unnamed. This combination of neural network works in a beautiful manner and it produces fascinating results.

4. Recurrent Neural Networks Extension

Over the years' researchers have developed more sophisticated types of RNNs to deal with some of the shortcomings of RNNs.

4.1 Bidirectional RNNs

This type of RNNs are based on the idea that the output at time t may not only depend on the previous elements in the sequence, but also future elements. For example, to predict a missing word in a sequence, we need to look at both the left and the right context. Bidirectional RNNs are quite simple. They are just two RNNs stacked on top of each other. The output is then computed based on the hidden state of both RNNs.

4.2 Deep (Bidirectional) RNNs

This type of RNNs are similar to Bidirectional RNNs, only that there are multiple layers per time step. In practice this gives a higher learning capacity but a lot of training data is required.

4.3 Long-Short-Term-Memory (LSTM)

LSTM is a special kind of RNNs that are designed to learn long term dependencies i.e., to memorize the sequence of data. The memory in LSTMs are called *cells* that are connected through layers. The cells resemble a transport line (the upper line in each cell) that connects out of one module to another one conveying data from past and gathering them for the present one. Each memory cell contains gates which handle information flow into and out of the cell. Internally these cells decide what to keep in (and what to erase from) memory. They then combine the previous state, the current memory and the input. Hence the gates which are based on sigmoidal neural network layer, enable the cells to optionally let data pass through or disposed. Each sigmoid layer yields numbers between 0 and 1, depicting every segment of data ought to be let through in each cell. More precisely, an estimation of

0 value implies that “let nothing pass through”; whereas; an estimation of one indicates that “let everything pass through”. There are three types of gates in the LSTM unit with the aim of controlling the state of each cell:

- Forget Gate outputs a number between 0 and 1, where 1 shows “completely keep this”; whereas, 0 implies “completely ignore this”.
- Memory Gate chooses which new data need to be stored in the cell. First, a *sigmoid* layer, called the “input door layer” chooses which values will be modified. Next a *tanh* layer makes a vector of new candidate values that could be added to the state.
- Output Gate decides what will be yield out of each cell. The yielded value will be passed on the cell state along with the filtered and newly added data.

The structure of the LSTM unit is shown in figure 5.

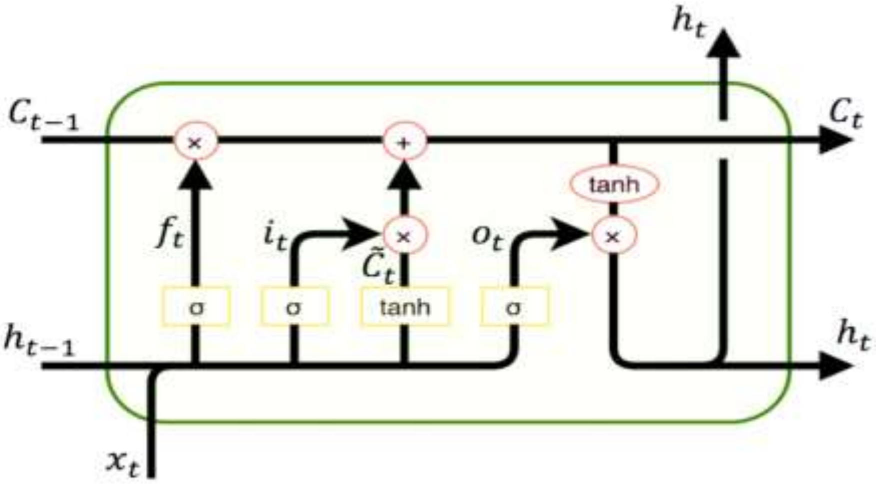


Fig. 5: LSTM unit

As seen from figure 5, Eq. v-vii, the LSTM unit gets the information from the previous state h_{t-1} and the input x_t, and uses the activation function (sigmoid) in the “input layer gate” to decide which part of the information to pass to the output and next LSTM unit.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad \text{(v)}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad \text{(vi)}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad \text{(vii)}$$

$$\widetilde{C_t} = tanh(W_C.[h_{t-1}, x_t] + b_C) \quad \text{(viii)}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \widetilde{C}_t \tag{ix}$$

$$h_t = o_t \otimes tanh(C_t) \tag{x}$$

Eq. v-vii describes the sigmoid function ($\sigma(x) = {}^{1}/_{1 + e^{-x}}$) where $W's$ and $b's$ are the parameters (weights and biases) for forget, input and output gates. f_t, i_t and o_t are forget, input and output gates respectively. In Eq. viii, the *tanh* layer creates the vector of new candidate values $\widetilde{C}_t$ that could be added to the cell state. LSTM unit has two kinds of hidden state: "slow" state C_t and a "fast" state h_t. The slow state C_t is updated by summing the multiplication the forget gate f_t by the previous cell state C_{t-1} and the multiplication the input gate i_t by the new candidate value $\widetilde{C}_t$. The h_t state is updated using the hyperbolic tangent function (*tanh*) of C_t state and o_t output gate. The calculation in the step is pretty much straightforward and eventually leads to the output. However, the outputs consist of only the outputs those were decided to be carry forwarded in the previous steps and not all the outputs at once. The main feature of LSTM unit is that its cell state accumulates activities over time. As derivatives of the error are summed over time, they do not vanish quickly. In this way, LSTMs can implement tasks over long term dependencies.

5. Conclusion

Artificial Intelligence (AI) has gained considerable prominence over the last decades fuelled by numerous applications in the field of image and speech recognition, automatic translation, modelling and forecasting, and many more areas. An artificial neural network is the field of artificial intelligence approach in which it tries to mimic the network of neurons that make up a human brain so that the computer will be able to learn things and make decisions in a human like manner. A recurrent neural networks (RNNs) is a class of neural networks that can use their internal state (memory) to process sequences of inputs and Long Short Term Memory (LSTM) networks are a kind of RNNs with its architecture for predicting sequence containing longer term patterns of unknown length, due to their ability to maintain long term memory.

6. Suggested Readings

Jha, G. K. and Sinha, K. (2014) Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India, *Neural Computing and Applications,* **24 (3-4)**, 563-571

Jha, G. K. and Sinha, K. (2013) Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System, *Agricultural Economics Research Review*, **26 (2),** 229-239

Mandic, D. & Chambers, J. (2001) Recurrent neural networks for prediction: Learning algorithms, architectures and stability, Wiley, ISBN 978-0-471-49517-8

Impact Assessment using Instrumental Variable and Propensity Score Matching Techniques

Anuja A R, Shivaswamy G P, K N Singh, Rajesh T and Harish Kumar HV
ICAR-IASRI, New Delhi

Research projects are designed with certain objectives. Projects might appear potentially promising before implementation yet fail to generate expected impacts or benefits. The obvious need for impact evaluation is to help researchers and policy makers decide whether projects or programs are generating intended effects; to promote accountability in the allocation of resources; and to fill gaps in understanding what works, what does not, and how measured changes in well-being are attributable to a particular project or policy intervention.

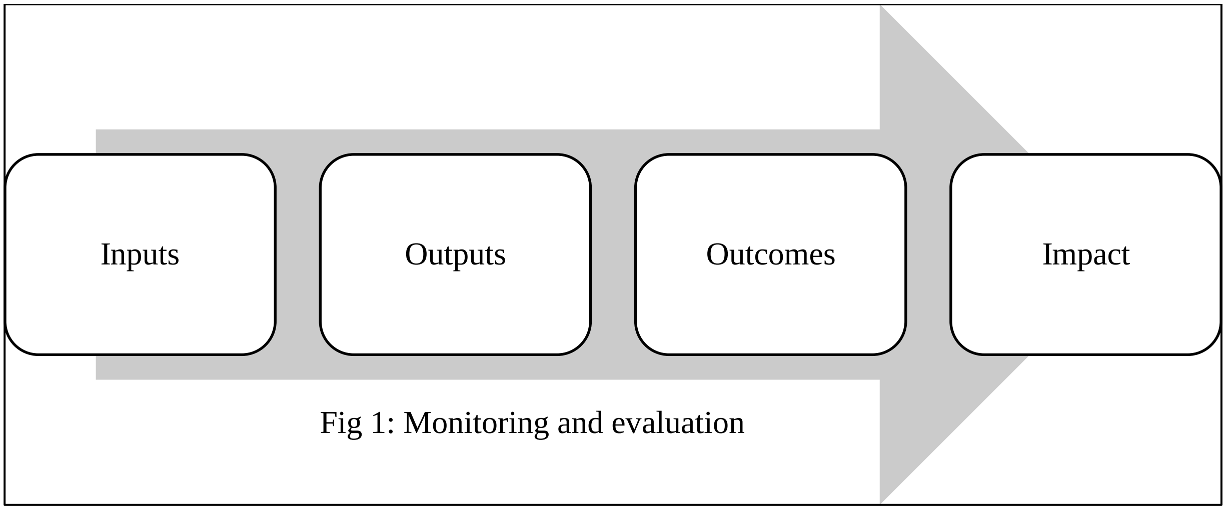


Fig 1: Monitoring and evaluation

What Are Inputs?

The resources used in a research are called inputs.

Example: Seeds, Fertilisers

What Are Outputs?

Outputs are the immediate results of the project/intervention. We usually describe outputs with numbers (example: Percentage increase in yield) and these are measurable and readily determined.

How Is an Outcome Different?

An outcome is an effect your research produces on the people or issues you serve or address. An outcome is a change that occurred **because** of your research. It is measurable and time-limited, although it may take a while to determine its full effect. Example: Farmers' income: Percentage increase in income.

Why Does Impact Matter?

Impacts are the **long-term or indirect effects** of your outcomes. Impacts are hard to measure since they may or may not happen. Example: Nutritional security

Impact evaluation

Impact can be defined as the "the attainment of development goals of the project or program, or rather the contributions to their attainment." Impact evaluation is an assessment of how the intervention being evaluated affects outcomes, whether these effects are intended or unintended. Impact evaluation spans qualitative and quantitative methods, as well as ex ante and ex post methods. Figure 1 portrays the different levels of impact evaluation.

When to do an impact evaluation

It is not feasible to conduct impact evaluations for all interventions. The following are examples of the types of intervention when impact evaluation would be useful:

- Innovative schemes
- Pilot programs which are due to be substantially scaled up
- Interventions for which there is scant solid evidence of impact in the given context
- A selection of other interventions across an agency's portfolio on an occasional basis

How to do an impact evaluation

Quantitative versus Qualitative Impact Assessments

Impact evaluation can be done through qualitative and quantitative techniques. Qualitative analysis seeks to measure potential impacts that the program may generate, the mechanisms of such impacts, and the extent of benefits to recipients from in-depth and group-based interviews. Whereas quantitative results can be generalizable, the qualitative results may not be.

Nonetheless, qualitative methods generate information that may be critical for understanding the mechanisms through which the program helps beneficiaries. Qualitative information such as understanding the local socio-cultural and institutional context, as well as program and participant details is, however, essential to a sound quantitative assessment. For example, qualitative information can help identify mechanisms through which projects might be having an impact; thereby aiding operational evaluation. But a qualitative assessment on its own cannot assess outcomes against relevant alternatives or *counterfactual outcomes*. That is, it cannot really indicate what might happen in the absence of the program. Quantitative analysis is also important in addressing potential statistical bias in program impacts. A mixture of qualitative and quantitative methods (a *mixed-methods approach*) might therefore be useful in gaining a comprehensive view of the program's effectiveness.

Quantitative Impact Assessment

Quantitative methods span ex-ante and ex-post approaches.

The **ex-ante design** determines the possible benefits or pitfalls of an intervention through simulation or economic models. This approach attempts to predict the outcomes of intended policy changes, given assumptions on individual behavior and markets. Ex-ante analysis can help in refining projects before they are implemented, as well as in forecasting the potential effects of programs in different economic environments.

The **Ex post impact evaluation**, in contrast, is based on actual data gathered either after program intervention or before and after program implementation. Ex post evaluations measure actual impacts accrued by the beneficiaries because of the program. These evaluations, however, sometimes miss the mechanisms underlying the program's impact on the population, which structural models aim to capture. These mechanisms can be very important in understanding program effectiveness (particularly in future settings).Ex post evaluations can also be much more costly than ex ante evaluations because they require collecting data on actual outcomes for participant and nonparticipant groups, as well as on other accompanying social and economic factors that may have determined the course of the intervention. An added cost in the ex post setting is the failure of the intervention, which might have been predicted through ex ante analysis. There are attempts to combine both these approaches.

Evaluation Design

The following are the key elements in designing an impact evaluation

- Deciding whether to proceed with the evaluation
- Identifying key evaluation questions
- The evaluation design should be embedded in the program theory
- The comparison group must serve as the basis for a credible counterfactual, addressing issues of selection bias (the comparison group is drawn from a different population than the treatment group) and contagion (the comparison group is affected by the intervention or a similar intervention by another agency).
- Findings should be triangulated
- The evaluation must be well contextualized

Impact evaluation: Major issues

The Problem of the Counterfactual

The main challenge of an impact evaluation is to determine what would have happened to the beneficiaries if the program had not existed. A beneficiary's outcome in the absence of the intervention would be its *counterfactual*. The problem of evaluation is that while the program's impact (independent of other factors) can truly be assessed only by comparing actual and counterfactual outcomes, the counterfactual is not observed. So the challenge of an impact assessment is to create a convincing and reasonable comparison group for beneficiaries in light of this missing data. Two methods that can be used include

- Before-and-After Comparisons
- With-and-Without Comparisons

The Problem of Selection Bias

Without information on the counterfactual, the next best alternative is to compare outcomes of treated individuals or households with those of a comparison group that has not been treated. In doing so, one attempts to pick a comparison group that is very similar to the treated group, such that those who received treatment would have had outcomes similar to those in the comparison group in absence of treatment.

Different Evaluation Approaches to Ex Post Impact Evaluation

A number of different methods can be used in impact evaluation theory to address the fundamental question of the missing counterfactual. Each of these methods carries its own assumptions about the nature of potential selection bias in program targeting and participation, and the assumptions are crucial for developing the appropriate model to determine program impacts.

These methods include
1.Randomized evaluations
2. Matching methods, specifically propensity score matching (PSM)
3. Double-difference (DD) methods
4. Instrumental variable (IV) methods
5.Regression discontinuity (RD) design and pipeline methods
6. Distributional impacts
7. Structural and other modeling approaches

These methods vary by their underlying assumptions regarding how to resolve selection bias in estimating the program treatment effect.

Randomized evaluations involve a randomly allocated initiative across a sample of subjects (communities or individuals, for example); the progress of treatment and control subjects exhibiting similar preprogram characteristics is then tracked over time. Randomized experiments have the advantage of avoiding selection bias at the level of randomization.

In the absence of an experiment**, PSM methods** compare treatment effects across participant and matched nonparticipant units, with the matching conducted on a range of observed characteristics. PSM methods therefore assume that selection bias is based only on observed characteristics; they cannot account for unobserved factors affecting participation.

DD methods assume that unobserved selection is present and that it is time invariant—the treatment effect is determined by taking the difference in outcomes across treatment and control units before and after the program intervention. DD methods can be used in both experimental and non-experimental settings.

IV models can be used with cross-section or panel data and in the latter case allow for selection bias on unobserved characteristics to vary with time. In the IV approach, selection bias on unobserved characteristics is corrected by finding a variable (or instrument) that is correlated with participation but not correlated with unobserved characteristics affecting the outcome; this instrument is used to predict participation.

RD and pipeline methods are extensions of IV and experimental methods; they exploit exogenous program rules (such as eligibility requirements) to compare participants and nonparticipants in a close neighborhood around the eligibility cutoff. Pipeline methods, in particular, construct a comparison group from subjects who are eligible for the program but have not yet received it.

1. Instrumental Variable Technique

In general, research issues in the social sciences are casual. Impact assessment studies focus on the influence of treatment on outcome. For example, while assessing the impact of a welfare initiative on poverty reduction, the welfare program is the treatment and poverty reduction is the intended outcome. Here, allotting treatment randomly to the experimental units is not feasible. Estimation of a causal relationship under such circumstances is problematic as it is difficult to establish that the treatments are exogenous to the investigated system.

One of the basic assumptions of Ordinary Least Square (OLS) is that there is no correlation between independent variables and residuals. When the predictor variable X is correlated with the error term U, the estimation of the causal effect using observational data will be biased. The problem can be addressed by adding additional exogenous variables to the model. In social science, Instrumental Variable (IV) technique is helpful to estimate the causal effect when there exists endogeneity. The Wu-Hausman test can be used to check endogeneity of treatment variable. IV can be used to solve the problem of omitted variable bias and the classic errors-in-variables problem.

Endogeneity occurs when there exists a correlation between independent variables and the error term. Let us take an example to explain the situation. Suppose we want to assess the impact of years of schooling on the earning of individuals. We observe correlation between years of schooling and the outcome variable i.e. earnings of individuals. But this correlation not necessarily indicates a causal relationship. Suppose, there is some unobservable variable that influences the outcome here such as IQ of the individual. There is a possibility that a better IQ of the individual is positively influencing both the treatment (years of schooling) and outcome variables (earnings of the individual). Figure 1 depicts the situations where causal inference in observational studies will be valid. The instrumental variable technique is an important tool used in the impact assessment studies in agriculture.

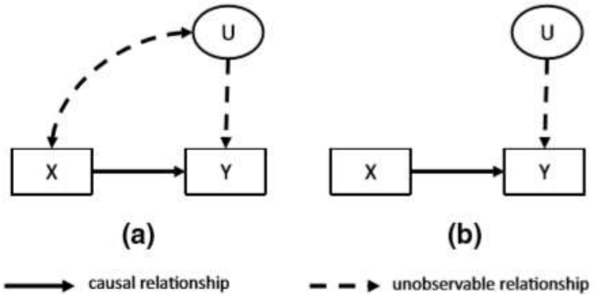


Fig 2 Examples of a situation where the modeling of causal relationships using observational data will be biased (a) and a situation where it will be valid (b) (Pokrope, 2016)

What are the instrumental variables?

Instrumental variable (IV) methods allow for endogeneity. An instrumental variable Z is an exogenous variable employed to assess the causal effect of variable X on Y (Figure 2).

A variable Z is an instrumental (relative to the pair (X, Y)) if

(i) Z is independent of all variables (including error terms) that have an influence on Y that is not mediated by X and

(ii) Z is not independent of X (Pearl 2000).

The first clause is referred to as the ‘exclusion’ and the second as the ‘relevance’.

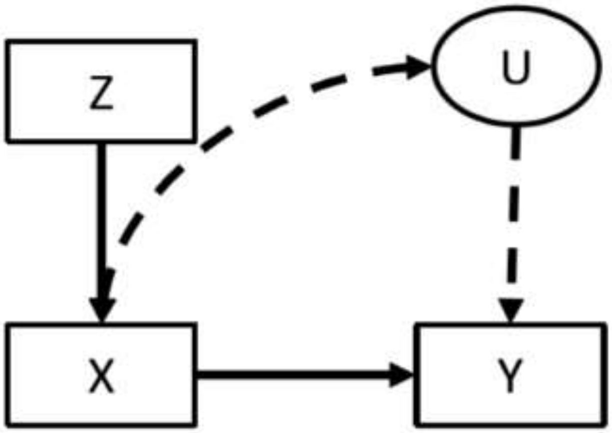


Fig.3 Situation where Z is a valid instrument (Pokrope, 2016)

Illustrating the application of instrumental variable technique in the agriculture

Birthal et al. (2015) employed IV technique to assess the impact of crop diversification on farm poverty in India. Unobserved features such a skill, motivation, etc. may lead to bias in the estimated coefficient. Using OLS regression to assess the impact may capture this unobserved heterogeneity and hence the estimates can suffer from bias. An instrumental variable was introduced into the model to mask unobserved heterogeneity at household level. As explained earlier, an ideal IV will not influence the outcome but will influence the treatment variable. In the study, the neighborhood effect based on geographical and social proximity was the IV. The logic of choosing the IV was that if the number of farmers growing high-value crops in the neighborhood is high it would positively influence the treatment variable i.e. area share of high-value crops. At the same time, the said IV would not affect the outcome variable of the model (farm poverty).

Selection of the instrumental variable

The selection of IV is of at most importance for the proper estimation of the causal effect. Finding a suitable instrumental variable for a large-scale database is a difficult task. Knowledge, experience and thorough understanding of the research issue can guide the researcher in finding proper IV for a situation. Weak instruments may worsen the bias in estimation (Khandker et al., 2010). A value greater than 10 for the first stage F statistic indicates a strong instrument. This does not necessarily rule out a weak instrument issue.

Disadvantages of instrumental variable

There are many challenges associated with the application of IV variables in impact assessment. The very difficulty in finding a suitable IV following all the assumptions is a major challenge. The poor performance of IV in small samples is another issue (Baum, 2008). The strength of the IV determines the precision. In comparison with the OLS estimates, IV estimates suffer from severe precision loss, if the instrument is weak. IV approaches are not immune from selection bias and the issue can be addressed by using the inverse probability of selection weights (Canan et al., 2017)

IV technique using Two-Stage Least Squares (2SLS) regression

In the OLS regression, there is a basic assumption that all independent variables are uncorrelated with the error term. Two-Stage least squares (2SLS) regression analysis is employed when there exists problem of endogeneity (Gujarati et al., 2012)

Problematic causal variable: This is the independent variable that is correlated with the error term or it is the variable that is influenced by other variables in the model. This endogenous causal variable is replaced with an instrumental variable in the first stage of the analysis.

Instrument variable: An instrumental variable is a new variable used in 2SLS to account for unobserved behavior between variables.

Estimation stages

First stage: A new variable is created using the instrument variable

Second stage: Instead of actual values of the problematic predictors, estimated values from the earlier stage is used in an OLS model to estimate the impact of the treatment variable

First stage regression:-

$$\boldsymbol{x_i = I\alpha + Z\,v + \delta_i} \qquad (1)$$

x_i – Vector of the endogenous variable i (where i = 1,…, N)

I- Matrix for Instrumental variables

Z- Matrix of the covariates

δ_i- Error term

The role of the instrumental variables finishes at the first stage of 2SLS. Covariates are included in the first stage of the estimation to ensure that there is no direct influence of IV on the outcome. More than one IV can be employed in the first stage considering the appropriateness of the variables.

Second stage regression: -

$$\boldsymbol{y} = \hat{\boldsymbol{x}}_{\boldsymbol{\iota}} \beta_i + \boldsymbol{Z\beta} + \boldsymbol{e} \qquad (2)$$

y- Vector of the outcome variable

$\hat{x}_\iota$- Vector of predicted values of *x* based on first stage regression

β_i - Parameter estimate of the causal effect of X on Y

Z- Matrix of the covariates

β - Vector of slope parameters for the covariates from Z

e - Error term.

Interpretation

The IV estimates indicate the local average treatment effect (LATE) instead of the average treatment effect (ATE). The ATE is the expected average effect of the treatment on outcome. The LATE provides information about the units that are likely to get the treatment if it is in the treatment group, but otherwise not take the treatment. The estimated LATE can be generalized for the population if there is no striking difference between the individuals influenced by the instrument and the population (Pokrope, 2016).

ILLUSTRATATION

Suppose we want to study the impact of having health insurance on medical expenses. In the given example, the dependent variable is ‘medical expenses’ (y_1), the endogenous regressor is ‘having health insurance’ (y_2) and exogenous regressors are illness, age, and income (x_1) of the individuals. In this example, social security income (ssi) ratio of the individual is used as an instrument (x_2). The IV represents variables assumed to affect ‘the choice of having health insurance or not’ but to have no direct effect on the outcome i.e. medical expenses. Table 1 indicates the sample data.

Table 1: Sample data

Number	Medical expenses	Health insurance	Age	Female	Income	Illnesses	ssi ratio
1	595	1	74	1	95	0	0.15
2	1783	1	73	0	36	3	0.40
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
n-1	720	0	69	1	29	1	0.15
n	809	1	90	1	21	1	0.36

Note: The data used in the illustrative example is a modified data from Katchova, A. (2013). Instrumental Variables in STATA. https://sites.google.com/site/econometricsacademy/econometrics-models/instrumental-variables.

OLS regression in STATA: -

First, define the dependent variable, independent variables, endogenous variable and instrumental variable. Command used for OLS regression in STATA – ‘***regress’***. Here the dependent variable is medical expenses (y_1). The endogenous regressor is ‘having health insurance’ (y_2) and exogenous regressors are illness, age, and income (x_1) of the individuals. Table 2 illustrates the results of OLS regression. The results indicate that for individuals with health insurance, the medical expenses are 7.5% higher than those for individuals without health insurance.

Command: regress y_1 y_2 x_{1list}

Table 2: Result of OLS regression

y_1: log of medical expenses	Coef.	SE	t	P>t	[95% Conf. Interval]	
Health insurance (y_2)	**0.075***	0.026	2.880	0.004	0.024	0.126
Illnesses (x_1)	0.441*	0.010	46.040	0.000	0.422	0.459
Age (x_1)	-0.003	0.002	-1.380	0.167	-0.006	0.001
Log of income (x_1)	0.017	0.014	1.250	0.211	-0.010	0.044
Constant	5.780*	0.151	38.310	0.000	5.484	6.076

* $p < 0.01$

2SLS estimation: - Command used for 2SLS regression using IV in STATA: ‘***ivregress’***

Command: ivregress 2sls y_1 (y_2= x_2) x_{1list}

Table 3: Result of 2 SLS estimation

y_1: log of medical expenses	Coef.	SE	t	P>t	[95% Conf. Interval]	
Health insurance (y_2)	**-0.852***	0.198	-4.300	0.000	-1.241	-0.463
Illnesses (x_1)	0.449*	0.010	43.590	0.000	0.428	0.469
Age (x_1)	-0.012*	0.003	-4.230	0.000	-0.017	-0.006
Log of income (x_1)	0.098*	0.022	4.350	0.000	0.054	0.142
SS incomer ratio (instrument x_2)	-					
Constant	6.590*	0.235	28.090	0.000	6.130	7.050

* $p < 0.01$

X_{1list} – Indicates list of exogenous variables

Table 3 explains the results of 2SLS with IV model. After instrumentation, for individuals with health insurance, their medical expenses are 85.2% lower than those for individuals without health insurance. It is evident from the results that the 2SLS coefficient turned out quite different from the OLS coefficient.

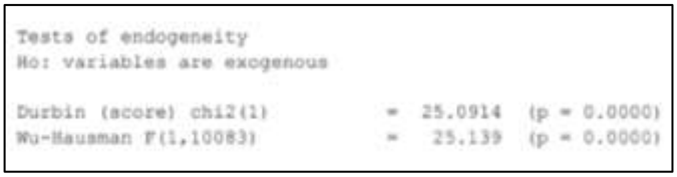
The following tests can be employed to ascertain the strength and suitability of the instruments.

Durbin-Wu-Hausman test for endogeneity

The endogeneity in the model can be tested using the Durbin-Wu-Hausman test for endogeneity. The Null hypothesis of the Durbin-Wu-Hausman test is that the independent variables are exogenous in nature. Rejection of null-hypothesis indicates the presence of endogeneity. The presence of endogeneity necessitates the usage of IV approach.

In the given example *test for endogeneity* was performed using the following command in STATA.

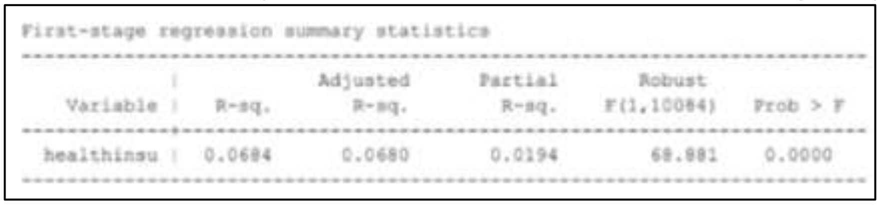
```
quietly ivregress 2sls y1 (y2= x2) x1list, first
estat endogenous
quietly regress  y2 x2 x1list
quietly predict vhat, resid
quietly regress  y1 y2 x1list vhat
testvhat
```



```
Tests of endogeneity
Ho: variables are exogenous

Durbin (score) chi2(1)          =  25.0914  (p = 0.0000)
Wu-Hausman F(1,10083)           =   25.139  (p = 0.0000)
```

The rejection of null hypothesis confirmed the presence of endogeneity.



```
First-stage regression summary statistics
--------------------------------------------------------------------------
             |            Adjusted      Partial       Robust
    Variable |   R-sq.       R-sq.        R-sq.     F(1,10084)   Prob > F
-------------+------------------------------------------------------------
  healthinsu |  0.0684      0.0680       0.0194        68.881    0.0000
--------------------------------------------------------------------------
```

Correlation

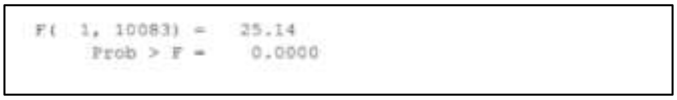
The correlation between ‘having health insurance’ (endogenous variable) and ssi (IV) was tested and there was a negative correlation of -0.21. Here the correlation is weak and this may lead to biased estimates.

Weak instrument test -F statistics

As a thumb rule, if the value of F statistics of the model is greater than 10, instruments are not weak. Following commands were used to estimate the F statistics.

quietly ivregress 2sls y_1 (y_2= x_2) x_{1list}, vce (robust)
estat first stage, forcenonrobust

As the value is 69 (which is greater than 10 as per thumb rule), the given instrument is not weak.



```
 F(  1, 10083) =   25.14
       Prob > F =    0.0000
```

Validity of multiple instruments.

The test for over-identifying restriction can be used to check the validity of multiple instruments. In the given example we have employed a single instrument.

2.Propensity Score Matching Technique

Propensity score matching (PSM) constructs a statistical comparison group that is based on a model of the probability of participating in the treatment, using observed characteristics. Participants are then matched on the basis of this probability, or propensity score, to nonparticipants. The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups. The validity of PSM depends on two conditions: (a) conditional Independence (namely, that unobserved factors do not affect participation) and (b) sizable common support or overlap in propensity scores across the participant and nonparticipant samples. Different approaches are used to match participants and nonparticipants on the basis of the propensity score. They include nearest-neighbor (NN) matching, caliper and radius matching, stratification and interval matching, and kernel matching and local linear matching (LLM). Regression-based methods on the sample of participants and nonparticipants, using the propensity score as weights, can lead to more efficient estimates.

On its own, PSM is a useful approach when only observed characteristics are believed to affect program participation. Whether this belief is actually the case depends on the unique features of the program itself, in terms of targeting as well as individual take up of the program. Assuming selection on observed characteristics is sufficiently strong to determine program participation, baseline data on a wide range of preprogram characteristics will allow the probability of participation based on observed characteristics to be specified more precisely. Some tests can be conducted to assess the degree of selection bias or participation on unobserved characteristics.

Given concerns with the implementation of randomized evaluations, the approach is still a perfect impact evaluation method in theory. Thus, when a treatment cannot be randomized, the next best thing to do is to try to mimic randomization—that is, try to have an observational analogue of a randomized experiment. With matching methods, one tries to develop a counterfactual or control group that is as similar to the treatment group as possible in terms of observed characteristics. The idea is to find, from a large group of nonparticipants, individuals who are observationally similar to participants in terms of characteristics not affected by the program. Each participant is matched with an observationally similar nonparticipant, and then the average difference in outcomes across the two groups is compared to get the program treatment effect. If one assumes that differences in participation are based solely on differences in observed characteristics, and if enough nonparticipants are available to match with participants, the corresponding treatment effect can be measured even if treatment is not random. The problem is to credibly identify groups that look alike. Identification is a problem because even if households are matched along a vector, X, of different characteristics, one would rarely find two households that are exactly similar to each other in terms of many characteristics. Because many possible characteristics exist, a common way of matching households is propensity score matching (PSM). In this technique, each participant is matched to a nonparticipant on the basis of a single propensity score, reflecting the probability of participating conditional on their different observed characteristics X.

What Does PSM Do?

PSM constructs a statistical comparison group by modeling the probability of participating in the program on the basis of observed characteristics unaffected by the program. Participants are then matched on the basis of this probability, or propensity score, to nonparticipants, using different methods outlined later in the chapter. The average Treatment effect of the program is then calculated as the mean difference in outcomes across these two groups. On its own, PSM is useful when only observed Characteristics are believed to affect program participation. This assumption hinges on the rules governing the targeting of the program, as well as any factors driving self-selection of individuals or households into the program. Ideally, if available, pre-program baseline data on participants and nonparticipants can be used to calculate the propensity score and to match the two groups on the basis of the propensity score. Selection on observed characteristics can also help in designing multi-wave experiments. Hahn, Hirano, and Karlan (2008) show that available data on covariates for individuals targeted by an experiment, say in the first stage of a two-stage intervention, can be used to choose a treatment assignment rule for the second stage—conditioned on observed characteristics. This equates to choosing the propensity score in the second stage and allows more efficient estimation of causal effects.

PSM Method in theory

The PSM approach tries to capture the effects of different observed covariates X on participation in a single propensity score or index. Then, outcomes of participating and nonparticipating households with similar propensity scores are compared to obtain the program effect. Households for which no match is found are dropped because no basis exists for comparison. PSM constructs a statistical comparison group that is based on a model of the probability of participating in the treatment T conditional on observed characteristics X, or the propensity score: P(X) = Pr(T = 1|X). The necessary assumptions for Identification of the program effect are (a) conditional independence and (b) presence of a common support. The treatment effect of the program using these methods can either be represented as the average treatment effect (ATE) or the treatment effect on the treated (TOT).

Assumption of conditional independence

Conditional independence states that given a set of observable covariates X that are not affected by treatment, potential outcomes Y are independent of treatment assignment T. If Y_i^T represent outcomes for participants and Y_i^C outcomes for nonparticipants, conditional independence implies

$$(Y_i^T, Y_i^C) \perp| T_i|X_i$$

This assumption is also called un-confoundedness, and it implies that uptake of the program is based entirely on observed characteristics. To estimate the TOT as opposed to the ATE, a weaker assumption is needed:

$$Y_i^C \perp| T_i|X_i$$

Conditional independence is a strong assumption and is not a directly testable criterion; it depends on specific features of the program itself. If unobserved characteristics determine program participation, conditional independence will be violated, and PSM is not an appropriate method. Having a rich set of preprogram data will help support the conditional independence assumption by allowing one to control for as many observed characteristics as might be affecting program participation.

Assumptions of common support

A second assumption is the common support or overlap condition: $0 < P(T_i = 1|X_i)$. This condition ensures that treatment observations have comparison observations nearby in the propensity score distribution. Specifically, the effectiveness of PSM also depends on having a large and roughly equal number of participant and nonparticipant observations so that a substantial region of common support can be found. For estimating the TOT, this assumption can be relaxed to $P (T_i= 1|X_i) < 1$

Treatment units will therefore have to be similar to non-treatment units in terms of I observed characteristics unaffected by participation; thus, some non-treatment units may have to be dropped to ensure comparability. However, sometimes a nonrandom subset of the treatment sample may have to be dropped if similar comparison units do not exist. This situation is more problematic because it creates a possible sampling bias in the treatment effect. Examining the characteristics of dropped units may be useful in interpreting potential bias in the estimated treatment effects. Treatment observations with weak common support can be dropped out. Only in the area of common support can inferences be made about causality, as reflected in figure 1. Figure 2 reflects a scenario where the common support is weak.

If conditional independence holds, and if there is a sizable overlap in P(X) across participants and nonparticipants, the PSM estimator for the TOT can be specified as the mean difference in Y over the common support, weighting the comparison units by the propensity score distribution of participants. A typical cross-section estimator can be specified as follows:

$$TOT_{PSM} = E_{P(X) \mid T = 1} \{E[Y^T| T=1, P(X)]-E[Y^C |T=0, P(X)]\}$$

Application of PSM method

To calculate the program treatment effect, one must first calculate the propensity score P(X) on the basis all observed covariates X that jointly affect participation and the outcome of interest. The aim of matching is to find the closest comparison group from a sample of nonparticipants to the sample of program participants. “Closest” is measured in terms of observable characteristics not affected by program participation.

First, the samples of participants and nonparticipants should be pooled, and then participation T should be estimated on all the observed covariates X in the data that are likely to determine participation. When one is interested only in comparing outcomes for those participating (T = 1) with those not participating (T = 0), this estimate can be constructed from a probit or logit model of program participation.

After the participation equation is estimated, the predicted values of T from the participation equation can be derived. The predicted outcome represents the estimated probability of participation or propensity score. Every sampled participant and non- participant will have an estimated propensity score, $\hat{P}(X|T = 1) = \hat{P}(X)$. Note that the participation equation is not a determinants model, so estimation outputs such as t-statistics and the adjusted R^2 are not very informative and may be misleading. For this stage of PSM, causality is not of as much interest as the correlation of X with T.

As for the relevant covariates X, PSM will be biased if covariates that determine participation are not included in the participation equation for other reasons. These reasons could include, for example, poor-quality data or poor understanding of the local context in which the program is being introduced. As a result, limited guidance exists on how to select X variables using statistical tests, because the observed characteristics that are more likely to determine participation are likely to be data driven and context specific. Bias in PSM program estimates can be low, given three broad provisions. First, if possible, the same survey instrument or source of data should be used for participants and non- participants. Using the same data source helps ensure that the observed characteristics entering the logit or probit model of participation are measured similarly across the two groups and thereby reflect the same concepts. Second, a representative sample survey of eligible nonparticipants as well as participants can greatly improve the precision of the propensity score. Also, the larger the sample of eligible nonparticipants is, the better matching will be facilitated. If the two samples come from different surveys, then they should be highly comparable surveys (same questionnaire, same interviewers or interviewer training, same survey period, and so on). A related point is that participants and nonparticipants should be facing the same economic incentives that might drive choices such as program participation. One could account for this factor by choosing participants and nonparticipants from the same geographic area.

Nevertheless, including too many X variables in the participation equation should also be avoided; over specification of the model can result in higher standard errors for the estimated propensity score $\hat{P}(X)$ and may also result in perfectly predicting participation for many households ($\hat{P}(X) = 1$). In the latter case, such observations would drop out of the common support (as discussed later). As mentioned previously, determining participation is less of an issue in the participating equation than obtaining a distribution of participation probabilities.

Defining the region of common support and balancing tests

Next, the region of common support needs to be defined where distributions of the propensity score for treatment and comparison group overlap. As mentioned earlier, some of the nonparticipant observations may have to be dropped because they fall outside the common support. Sampling bias may still occur, however, if the dropped nonparticipant observations are systematically different in terms of observed characteristics from the retained nonparticipant sample; these differences should be monitored carefully to help interpret the treatment effect.

Balancing tests can also be conducted to check whether, within each quantile of the propensity score distribution, the average propensity score and mean of X are the same. For PSM to work, the treatment and comparison groups must be balanced in that similar propensity scores are based on similar observed X. Although a treated group and its matched non-treated comparator might have the same propensity scores, they are not necessarily observationally similar if misspecification exists in the participation equation. The distributions of the treated group and the comparator must be similar, which is what balance implies. Formally, one needs to check if $\hat{P}(X \mid T = 1) = \hat{P}(X \mid T = 0)$.

Matching participants to nonparticipants

Different matching criteria can be used to assign participants to non-participants on the basis of the propensity score. Doing so entails calculating a weight for each matched participant-nonparticipant set. As discussed below, the choice of a particular matching technique may therefore affect the resulting program estimate through the weights assigned:

Nearest-neighbor matching: One of the most frequently used matching techniques is nearest neighbor (NN) matching, where each treatment unit is matched to the comparison unit with the closest propensity score. One can also choose n nearest neighbors and do matching (usually n = 5 is used). Matching can be done with or without replacement. Matching with replacement, for example, means that the same non- participant can be used as a match for different participants.

Caliper or radius matching: One problem with NN matching is that the difference in propensity scores for a participant and its closest nonparticipant neighbor may still be very high. This situation results in poor matches and can be avoided by imposing a threshold or "tolerance" on the maximum propensity score distance (caliper). This procedure therefore involves matching with replacement, only among propensity scores within a certain range. A higher number of dropped non- participants is likely, however, potentially increasing the chance of sampling bias.

Stratification or interval matching: This procedure partitions the common support into different strata (or intervals) and calculates the program's impact within each interval. Specifically, within each interval, the program effect is the mean difference in outcomes between treated and control observations. A weighted average of these interval impact estimates yields the overall program impact, taking the share of participants in each interval as the weights.

Kernel and local linear matching: One risk with the methods just described is that only a small subset of nonparticipants will ultimately satisfy the criteria to fall within the common support and thus construct the counterfactual outcome. Nonparametric matching estimators such as kernel matching use a weighted average of all nonparticipants to construct the counterfactual match for each participant.

Calculating the average treatment effect

As discussed previously, if conditional independence and a sizable overlap in propensity scores between participants and matched nonparticipants can be assumed, the PSM average treatment effect is equal to the mean difference in outcomes over the common support, weighting the comparison units by the propensity score distribution of participants. To understand the potential observed mechanisms driving the estimated program effect, one can examine the treatment impact across different observable characteristics, such as position in the sample distribution of income, age, and so on.

Estimating Standard Errors with PSM: Use of the Bootstrap

Compared to traditional regression methods, the estimated variance of the treatment effect in PSM should include the variance attributable to the derivation of the propensity score, the determination of the common support, and (if matching is done without replacement) the order in which treated individuals are matched. Failing to account for this additional variation beyond the normal sampling variation will cause the standard errors to be estimated incorrectly.

One solution is to use bootstrapping, where repeated samples are drawn from the original sample, and properties of the estimates (such as standard error and bias) are estimated with each sample. Each bootstrap sample estimate includes the first steps of the estimation that derive the propensity score, common support, and so on. Formal justification for boot- strap estimators is limited; however, because the estimators are asymptotically linear, bootstrapping will likely lead to valid standard errors and confidence intervals.

Advantages and disadvantages of PSM

The main advantage and drawback of PSM relies on the degree to which observed characteristics drive program participation. If selection bias from unobserved characteristics is likely to be negligible, then PSM may provide a good comparison with randomized estimates. To the degree participation variables are incomplete, the PSM results can be suspect. This condition is not a directly testable criteria; it requires careful examination of the factors driving program participation.

Another advantage of PSM is that it does not necessarily require a baseline or panel survey, although in the resulting cross-section, the observed covariates entering the logit model for the propensity score would have to satisfy the conditional independence assumption by reflecting observed characteristics X that are not affected by participation. A preprogram baseline is more helpful in this regard, because it covers observed X variables that are independent of treatment status.

REFERENCES

Baum, C. F. (2008), Using Instrumental Variables Techniques in Economics and Finance. Boston College and DIW Berlin. German Stata Users Group Meeting, Berlin, June 2008. Online available at https://www.stata.com/meeting/germany08/Baum.DESUG8621.beamer.pdf

Birthal, P. S., Roy, D. and Negi, D. S. (2015), Assessing the Impact of Crop Diversification on Farm Poverty in India. World Development, Elsevier, 72(C), 70-92.

Canan, C., Lesko, C., & Lau, B. (2017), Instrumental Variable Analyses and Selection Bias. *Epidemiology (Cambridge, Mass.)*, *28*(3), 396–398. doi:10.1097/EDE.0000000000000639

Gujarati, D.N., Porter, D.C. and Gunasekar, S. (2012), Basic Econometrics. McGraw Hill Education (India) Private Limited.

Hahn, J. Hirano, K.and Karlan, D. (2008) Adaptive Experimental Design Using the Propensity Score. Journal of Business and Economic Statistics 29(1):96-108

Katchova, A. (2013), Instrumental Variables in STATA. Online available at https://sites.google.com/site/econometricsacademy/econometrics-models/instrumental-variables.

Khandker, S.R., Koolwal, G.B. and Samad, H.A. (2010), World Bank: Handbook on Impact Evaluation: Quantitative Methods and Practices. World Bank.

Pearl, J. (2000), Causality: Models, Reasoning and Inference. New York: Cambridge University Press.

Pokrope, A. (2016), Introduction to Instrumental Variables and their Application to Large-Scale Assessment Data. Large-scale Assessments in Education 4:4 DOI 10.1186/s40536-016-0018-2.