תוכנית כנס האיגוד הישראלי לסטטיסטיקה - 28 מאי 2015

לשכה המרכזית לסטטיסקה

שעה	מיקום	כותרת		מרצה
		סדר היום של כנס האיגוד הישראלי לסטטיסטיקה 2015		
שעה	מיקום	כותרת		מרצה
08:00 – 08:30	מבואה	תליית פוסטרים		
08:30 – 09:00	מבואה	הרשמה		
09:00 – 10:20	מליאה	דברי ברכה Methodological Issues and Challenges in the Production of Official Statistics		דוד שטיינברג דני פפרמן
10:20 – 10:40	מליאה	אספת מליאה + הענקת פרסים		
10:40 – 11:00	מליאה	קדימון למושב פוסטרים		
11:00 – 11:40	מבואה	מושב פוסטרים + כיבוד		
		Calculation of Likelihood Ratios for Gunshot Residue Evidence - Statistical Aspects		נעמי קפלן דמרי
		Support Vector Machines for Current Data		יעל טרוויס לומר
		Analytical and Numerical Investigation of Stochastic Ricker Processes for Population Dynamics		תמר גדריך
		The dendextend R Package for the Visualization and Comparison of Dendrograms		טל גלילי
		Early Detection of Long Term Evaluation Criteria in Online Controlled Experiments		לירון גת כחלון \ יוני שחמרוט
		Randomized Allocations of Subjects to Experimental Treatment Groups		יונתן יפה נוף
		Market Depth of Israeli Corporate Bonds: Comparison to OECD Member States		אלרואי חדד
		Sparse Clustering of Noisy Signals in the Wavelet Domain		תום הוף ואבישי וגנר
11:40 – 13:20	מושב 1 EMR-IBS	סטטיסטיקה בפיתוח תרופות		אירגון ויו"ר: ענת סאקוב
		התמודדות עם ערכים חסרים ברוח ההנחיות העדכניות של הFDA		ניר שרון
		שיקולים סטטיסטיים בניסויים קליניים פאזה 1		הדס ברקאי
		שיקולים סטטיסטיים ב- Adaptive Designs		ענת סאקוב
		Learning DNA mutation patterns in error-prone PCR experiments		יובל נוב
	מושב 2	"סטטיסטיקה בעולם הדיגיטלי"		אירגון ויו"ר: אפרים גולדין
		Statistics in the Big Data world		דרור יוסף
		Measuring Risk in the world of online social trading		ניר שמולביץ
		Statistical Challenges in the World of Web and Mobile Applications		יוני שחמרוט
				מרצה 4

שעה	מיקום	כותרת	מרצה
סדר היום של כנס האיגוד הישראלי לסטטיסטיקה 2015			
13:20 – 14:30	מסעדות	ארוחת צהריים	
14:30 – 16:10	מושב 3 EMR-IBS	נתונים חסרים	אירגון ויו"ר: חבי מורד
		Accounting for Measurement Error in Genomic Data and Misclassification of Subtypes in the Analysis of Heterogenous Tumor Data	דניאל נבו
		Estimating and Testing Interactions using Multiple imputation,when Explanatory Variables Are Subject to Measuremnet Error	חבי מורד
		דה-קונוולוציה וטיפול באי-השבה	איתן גרינשטיין
		Dynamic Model to Study the Association Between Two Longitudinal Measures in the Presence of Survival	דיקלה גבע
	מושב 4	סטטיסטיקה רשמית	יו"ר: צחי מקובקי איגרון: לואיזה ברוק
		Estimation of Dwelling Values in Israeli Arab Sector	יורי גובמן
		Modelling the Nonlinear Statistics of Cities in Israel	דוד בורג
		בקרת איכות דו-רמתית על ציונים לפי משתנים דמוגרפיים	דביר קלפר
		הערכת דיוק המדד הוצאה למזון : הערכת דיוקו של מדד מחירי המזון למרכיביו: בחינה אמפירית באמצעות נתוני סורק	דורון סייג
16:10 – 16:30	מבואה	הפסקת קפה	
16:30 – 16:40	מליאה	הענקת פרס לפוסטר הנבחר	
16:40 – 17:40	מליאה	The Promise and Perils of Big Data for Statistical Inference	סטיב פיינברג
17:40 – 18:40	מליאה	פנל: האתגרים האקנומטרים/סטטיסטים של כלכלה ומימון בני זמננו	יו"ר: יצחק מלכסון

Methodological Issues and Challenges in the Production of Official Statistics

Danny Pfeffermann

Danny Pfeffermann, Government Statistician of Israel
Professor, Hebrew University of Jerusalem, Israel
&
Southampton Statistical Sciences Research Institute, UK

Email addresses: msdanny@soton.ac.uk ; msdanny@cbs.gov.il

 The big advancement in technology, coupled with increased availability of 'big data', and yet increased demand for more accurate, more detailed and more timely official data with tightened budgets, places enormous challenges to producers of official statistics across the world. In this presentation I shall discuss some of the major challenges as I see them and in some cases offer ways of dealing with them. Examples include the potential use of big data; privacy and confidentiality; possible use of data obtained from web-panels; accounting for mode effects; and the integration of administrative data and small area estimation for future censuses. In the last part of my talk I shall confront the question of whether universities train their students to work at National Statistical Offices.

The Promise and Perils of Big Data for Statistical Inference

Stephen E. Fienberg

Department of Statistics, Heinz College,
&
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Claims for the importance of big data abound: "Analysis of big datasets can find new correlations, spot business trends, prevent diseases, combat crime, and replace traditional surveys and censuses." Are such claims warranted? In a number of traditional statistical domains, others have suggested that we substitute "big data" for more traditional forms of survey and experimental data. In this talk I will discuss the role of big data in a variety of different settings, describing some of the exciting new prospects and why I think traditional statistical ways of thinking about the collection and analysis of statistical data are more important than ever.

פנל: האתגרים האקנומטרים/סטטיסטים של כלכלה ומימון בני זמננו

Early Detection of Long Term Evaluation Criteria in Online Controlled Experiments

Liron Gat Kahalon[1] and Yoni Schamroth[1] and David Steinberg[2]

[1]Perion Networks,   [2]Tel-Aviv University

Controlled Experimentation has been universally adopted by the online world as an essential tool in aiding in the decision making process and has been widely recognized as a successful scientific method for establishing causality. Frequently referred to A/B testing or multivariate testing, controlled experiments provide a relatively straightforward method for quickly discovering the expected impacts of new features or strategies. One of the main challenges involved in setting up an experiment is deciding upon the OEC, or overall evaluation criteria. In this paper, we demonstrate the importance of choosing a metric that focuses on long term effects. Such metrics include measures such as life-span or lifetime value. We present an innovative methodology for early detection of lifetime differences between test groups. Finally we present motivating examples where failure to focus on the long term effect may result in an incorrect conclusion.

Sparse Clustering of Noisy Signals in the Wavelet Domain

Tom Hope, Avishai Wagner and Or Zuk

Department of Statistics, Hebrew University of Jerusalem, Jerusalem, Israel

We propose Sparse Wavelet Domain Clustering (SWDC) - a computationally efficient time-series clustering method that exploits sparse representation of signals in the wavelet domain, which applies dimensionality reduction and clustering simultaneously, via a 'built-in' shrinkage of the wavelets coefficients based on their contribution to the clustering information. We show that standard methods, based on signal smoothing and K-means, lose vital cluster information and yield poor performance when applied to clustering noisy signals. In contrast to standard shrinkage of individual signals, our method pools information from multiple instances simultaneously, thus increasing the signal-to-noise ratio, and in addition shrinks coefficients in a manner optimized for the clustering task at hand, by using the Sparse K-means algorithm on the wavelets coefficients. These two properties enable us to identify the global clustering structure, especially when clustering information is concentrated in a few coefficients. Simulations show that our method leads to significant improvements in accuracy over standard methods, with similar computational cost. We also propose an efficient translation-invariant clustering method, and demonstrate the method's ability to correctly group shifted signals belonging to the same cluster.

The Dendextend R Package For The Visualization And Comparison of Dendrograms

Tal Galili[1], Yoav Benjamini[1]

[1]*Department of Statistics and Operations Research,*
The Sackler Faculty of Exact Sciences, Tel Aviv University.

Email: *tal.galili@gmail.com*

In this poster I will introduce the **dendextend** package[1] which extends the palette of functions and methods for the `dendrogram` class.

A dendrogram is a tree diagram which is often used to visualize a hierarchical clustering of items. Dendrograms are used in many disciplines, ranging from Phylogenetic Trees in computational biology to Lexomic Trees in text analysis. Hierarchical clustering in *R* is commonly performed using the `hclust` function. When a more sophisticated visualization is desired, the `hclust` object is often coerced into a `dendrogram` object, which in turn is modified and plotted. While **base** *R* comes with several very useful methods for manipulating the `dendrogram` object (namely: `plot, print, [[, labels, as.hclust, cophenetic, reorder, cut, merge, rev,` and `str`), still - the current palette of functions leaves a lot to be desired.

The novel **dendextend** package offers functions and methods for the `dendrogram` object, allowing for easier manipulation of a dendrogram's shape, color and content through functions such as `rotate, prune, labels<-, labels_colors, cutree, color_branches,` and more. **dendextend** also provides the tools for comparing the similarity of two dendrograms to one another either graphically using a `tanglegram` plot, or statistically with association measures ranging from `cor_cophenetic` to `Bk_plot`, while enabling bootstrap and permutation tests for comparing the trees. Since tree structure often requires the use of recursion, which can be slow in *R*, some of the more computationally intensive aspects of the **dendextend** package can be handled with its sister package, **dendextendRcpp**[2], which overrides several basic functions (namely: `cut_lower_fun, heights_per_k.dendrogram, labels.dendrogram`), with their *C++* implementation.

Keywords: dendrogram, dendextend, hierarchical clustering, R

Market Depth of Israeli Corporate Bonds:
Comparison to OECD Member States

Elroey Hadad, Gitit Gur-Gershgoren and Haim Kedar-Levy

Ben-Gurion University

This research aims to study market depth of Israeli High Quality Corporate Bonds (HQCB), as compared with OECD member states. The motivation stems from International Accounting Standard 19 (IAS-19), which Israel adopted recently, whereby long-term employee benefits should be discounted by reference to market yields on local HQCB at the end of the reporting period. If there is no deep market in HQCB, IAS-19 requires that the discount rate be determined based on local government bond yields. The choice between the two rates might have a substantial impact on the entity's financial statements (For example, in Israel Electric the impact is about 4 billion Shekels, and in the big banks between 0.5-2 billions). However, IAS-19 does not provide a clear definition of the criteria by which one would define High Quality of Corporate bonds, and it is not specific about the particular definition of market depth. This study aims to develop a coherent and unified framework for defining which bonds may be considered High Quality, and how market depth can and should be measured. By answering these questions, we aim to provide a uniform and comprehensive methodology to explore whether a country can be classified as having a deep market, or not. Our first step has already been conducted, where we examined whether the Israeli HQCB market may be considered deep or not.

Market depth is often defined as a market in which traders' transactions have small impact on prices. To measure depth, we examined the relevance of various financial variables as predictors for a binary market-depth indicator. We applied Linear Discriminant Analysis (LDA) over recent in-sample data from 31 markets, obtained from the World Bank. The analysis focuses on in-sample performance from one to three years before and after the global financial crisis of 2008. Results show that the relative sizes of deposits in financial institutions as well as the private debt securities to GDP are robust predictors for financial market depth. Out-of-sample results for the Israeli market show discriminant scores within the range of the deep markets group with evaluated group membership above 99%, outperforming the classification results of well-developed markets such as Italy, Finland and Sweden. The findings of this research have been accepted by the Israeli Securities Authority, who declared that Israel has a deep market in HQCBs, and hence instructed on immediate implementation of IAS-19's requirements.

Analytical and numerical investigation of Stochastic Ricker Processes for population dynamics

Tamar Gadrich[1], Guy Katriely[2]

[1]*Department of Industrial Engineering and Management, ORT Braude College,Karmiel, Israel.*
[2]*Department of Mathematics, ORT Braude College, Karmiel, Israel.*

The Ricker model is a well-known discrete-time deterministic model used by theoretical ecologists to describe density-dependent population growth, and is one of the simplest population models displaying complex dynamical phenomena like periodicity and chaos. Though originally the Ricker model was introduced in a phenomenological way. In recent years researchers (e.g:, Brannstrom, Johansson, Sumpter, Royama) have been able to derive the Ricker model from basic principles using the `site-based' approach (i.e., local interactions between individuals inuence population dynamics). These derivations show that the Ricker model is obtained as a `large population' limit of a class of discrete-time Markov processes, which can be called `Ricker processes'. Studying these processes is interesting, since it allows us to explore the interaction between the complex dynamics of the deterministic limit and the stochastic effects which are important for relatively small populations. This is the purpose of the present study. We derive explicit expressions for the transition probabilities of Ricker processes, using `balls and bins' and generating function techniques, and allowing for an arbitrary reproductive distribution (distribution of the number of offspring per individual). These expressions are used, together with numerical computations, to explore the quasi-stationary distributions and the time to extinction of Ricker processes. The expressions for the transition probabilities can also be used for fitting these models to real-world data using the method of maximum likelihood.

Calculation of Likelihood Ratios for Gunshot Residue Evidence - Statistical Aspects

Naomi Kaplan Damary , Micha Mandel, Nadav Levin, Elad Izraeli

Hebrew University of Jerusalem

Recently published articles have proposed the use of likelihood ratios (LRs) in determining the evidential value of finding a given number of gunshot residue (GSR) particles on a suspect. LRs depend on the models assumed for the null hypothesis (the suspect was not involved in a shooting) and the alternative hypothesis (the suspect was involved). In this article, data from Cardinetti et al. (2006) are used to demonstrate the sensitivity of the LR to the model. It is shown that the Poisson model frequently assumed for GSR data is inappropriate for describing the distribution of particles found on subjects involved in a shooting, and that a Negative Binomial model fits the data much better. The statistical error arising from the fact that models are estimated based on small sampled data is discussed, as well as the importance of accounting for this error. We conclude that before LRs can be used in the field of GSR analysis, many more samples must be collected. Only with a large database, can statistical models be estimated accurately and LR's be treated as valid scienti_c measures.

Support Vector Machines For Current Status Data

Yael Travis-Lumer

Haifa University

Current status data is a data format where the time to event is restricted to knowledge of whether or not the failure time exceeds a random monitoring time. We develop a support vector machine learning method for current status data that estimates the failure time expectation as a function of the covariates. In order to obtain the support vector decision function, we minimize a regularized version of the empirical risk with respect to a data-dependent loss. We show that the decision function has a closed form. Using finite sample bounds and novel oracle inequalities, we prove that the obtained decision function converges to the true conditional expectation for a large family of probability measures and study the associated learning rates. Finally we present a simulation study that compares the performance of the proposed approach to current state of the art.

Randomized Allocations of Subjects to Experimental Treatment Groups

Jonathan Yefenof[1]

[1] QBI Enterprises.

Email: *jonathan.yefenof@mail.huji.ac.il*

In science, experiments that involve blindness and hence randomized, are those that allow reliability and validity of statistical estimates for treatment outcomes. In the statistical theory of experiment design, randomization involves stochastic allocation of experimental units across the treatment groups. The units in a trial are classified to two categories: 1) All units are available in the beginning of the trial (typically in in-vivo studies), 2) The units are recruited along trial duration (typically in clinical studies). We are proposing a method of randomization which is related to the scenario where all the units are aviable in the beginning.

We address the need to allocate subjects to treatment groups. It is desirable that the groups will be homogeneous, i.e. the average and variance of the measurements in each group should be similar. Assuming that the distribution of the observations has enough moments, we show that our method succeeds in preserving the homogeneity of the groups.

Keywords: Randomization, Blindness, In-Vivo, Homogeneity, Moments.

התמודדות עם ערכים חסרים ברוח ההנחיות העדכניות של הFDA

ניר שרון,

Quark Pharmaceuticals Inc.

בשנת 2010 רשות המזון והתרופות האמריקאית (ה-FDA) אימצה הנחיות לטיפול בערכים חסרים אשר מתבססות על שימוש בשיטות בעלות הצדקה תיאורטית לטיפול בערכים חסרים. עד 2010 ה-FDA איפשר טיפול בערכים חסרים בשיטות אד-הוק להן אין צידוק תיאורטי ובהן קשה להעריך את כיוון ההטיה ועוצמתה. בהרצאה זו יוצג ניתוח תוצאות של ניסוי קליני פאזה שנייה תוך טיפול בערכים חסרים לאור ההנחיות העדכניות. התוצאות יושוו לשיטות עבר מקובלות תוך התייחסות לדפוס הערכים החסרים, הסיבות להיווצרותם, בניית המודל לטיפול בערכים חסרים ובדיקות הרגישות למודל שנבחר.

שיקולים סטטיסטיים בניסויים קליניים פאזה 1

הדס ברקאי,

טבע תעשיות פרמצבטיות בע"מ

ניסויי פאזה 1 הם השלב הראשון בניסויים קליניים לפיתוח תרופות, ושזורים בתכנית הפיתוח של כל תרופה חדשה עד לאישורה. בהרצאה זו יוצגו מטרות ניסויי פאזה 1, מושגי יסוד בפרמקוקינטיקה (Pharmacokinetics) ופרמקודינמיקה (Pharmacodynamics) והשיקולים הסטטיסטיים בתכנון ניסויים מסוגים שונים. כמו כן יוצג המבחן הסטטיסטי הנדרש ע"י ה FDA- בניסויי ביואקוויוולנטיות (bioequivalence) והפרמטרים הדרושים לחישוב גודל המדגם.

שיקולים סטטיסטיים ב- Adaptive Designs

ענת סאקוב,

טבע תעשיות פרמצבטיות בע"מ

ניסויים קליניים מסווגים באופן מסורתי לשלוש פאזות שלכל אחת מטרות שונות. מתוך מטרה לקצר זמני פיתוח, אנו רואים בשנים האחרונות שימוש ב- adaptive design שמנסים לענות על מטרות שונות בניסוי אחד. בהרצאה זו אתן סקירה מה הם adaptive designs ומה הם האתגרים והשיקולים הסטטיסטיים בניסויים כאלו.

Learning DNA mutation patterns in error-prone PCR experiments

Yuval Nov,

Haifa University

 Error-prone Polymerase Chain Reaction (epPCR) is a widely used protein engineering technique, whereby a gene is subject to several rounds of mutation-prone replication. Statistical analysis of experimental data reveals that the likelihood of mutation arising in the process fluctuates greatly but consistently across the gene. We study how local DNA patterns influence the likelihood of mutation, while taking into account statistical identifiability issues arising from the complementary nature of the replication process. We also discuss how these patterns may be exploited to improve the evolvability of genes in epPCR experiments.

Joint work with Dan Tawfik and Liat Rockah-Shmuel from the Weizmann Institute of Science.

Statistics in the Big Data world

Dror Yosef

CDO STARTAPP

Machine learning, big data and data science are hot topics (or buzz words) in the high tech community nowadays.
 In this session I will share my point of view about the differences between statistics and machine learning (if any) , and how they are used today in the computational evolving area.

Measuring Risk in the world of online social trading

Nir Szmulewicz

Director of BI, eToro

In social trading, traders can allocate a portion of their portfolio and copy investment activity of some other trader, as they were asset, Indices or CFD of FX Pairs this kind of trading creates some Ethical questions and methodologies complications. One of the key issues is the Risk Award calculation of the underlying asset, which is now a flash and bones human trader.
 I will show how by using some basic statistical fundamentals we resolved the risk measurement question, and reveal the various applications eToro had implemented based on this new metric

Statistical Challenges in the World of Web and Mobile Applications

Yoni Schamroth
Perion Networks

With the recent explosion of data availability vast amounts of knowledge lie dormant waiting to be plucked. This so called 'data revolution' has found companies desperately seeking those experts skilled in analyzing, making sense of and exploiting this overwhelming cascade of information. We present some of the challenging areas data scientists are required to contend with in the domain of web and mobile applications. These areas include real-time bidding systems, recommendation engines, life-time value calculations and A/B testing.

תקציר 4

Accounting for Measurement Error in Genomic Data and Misclassification of Subtypes in the Analysis of Heterogenous Tumor Data

Daniel Nevo

Hebrew University

A common paradigm in dealing with heterogeneity across tumors in cancer analysis is to cluster the tumors into subtypes using biomarker data on the tumor and then to analyze each of the clusters separately. A more specific target is to investigate the association between risk factors and specific subtypes. This task is usually carried out by two steps – clustering and risk factor assessment. However, two sources of measurement error arise in these problems. The first is the error in the biomarker measurements. The second is the misclassification error when assigning observations to clusters. We consider the case with a specified set of relevant markers and propose a unified single-likelihood approach for normally distributed biomarkers. As an alternative, we consider a two-step procedure with the tumor type misclassification error taken into account in the second-step risk factor analysis. We describe our method for multinomial data and also for survival analysis data using a modified version of the Cox model.

דה-קונוולוציה וטיפול באי-השבה

Eitan Greenstein

Central Bureau of Statistics

We consider the problem of estimating the sum of n latent-variables/parameters, based on corresponding n observations. The relation to the problem of estimation under censoring or truncation is explained, and in particular treatment of non-response. The relation to the problem of de-convolution and more generally identifying mixing Distributions is also explained. We briefly consider the computation involved and show some numerical results. Based on a joint work with Theodor Itskov.

Estimating and Testing Interactions Using Multiple Imputation, When Explanatory Variables are Subject to Measurement Error

Havi Murad[1], Laurence S. Freedman[1], David Zucker[2]

[1]*Biostatistics Unit, Gertner Institute for Epidemiology and Health Policy Research, ISRAEL*
[2]*Department of statistics, Hebrew University of Jerusalem, ISRAEL*

Email: *HaviM@gertner.health.gov.il*

Assessing interactions in linear regression models when covariates are subject to measurement error (ME) is complex since the error in the interaction term is a product of errors. In previous work we have described a method based on Regression Calibration (RC), efficient Normal Based RC (NBRC), for estimating interaction coefficients of normally distributed covariates having classical1 and non-classical2 normally distributed ME. This method yielded consistent estimators and standard errors for the interaction under the assumption of non-differential error, i.e., that the covariate errors are independent of Y, conditional on the X's.

In this presentation we focus on the use of multiple imputation (MI) for addressing this problem. We assume there are covariates W that are fully observed and measure the true covariates X, which are assumed to follow a multivariate normal distribution, with non-classical ME. The X's are unobserved for all subjects, and we impute them using MI via the MCMC algorithm. This involves random draws from the conditional distribution of the X's given all the observed variables using the accept-reject sampling method. We assume that there is a sub-sample of the participants that have unbiased normally distributed reference measurements R for X as well as W and Y. The imputations are performed for subjects not in the sub-study using random draws from the conditional density function of X on W and Y, and for subjects in the sub-study using random draws conditional on W, R and Y. The interaction term is passively imputed as the product of the imputed X's. This method is more general than efficient NBRC since it does not assume non-differential error.

We apply the method to data from the Observing Protein and Energy Nutrition study (OPEN), and examine the relationship of BMI with protein and energy intake in males. We use simulations to compare MI with efficient NBRC under non-differential and differential error.

Keywords: missing data, multiple imputation, measurement error, interaction, regression calibration

[1]Murad H and Freedman LS. Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. *Stat Med* 2007; **26:**4293-4310

[2]Murad H, Kipnis V, and Freedman LS. Estimating and testing interactions when explanatory variables are subject to non-classical measurement error. *SMMR* 2013; **DOI:** 10.1177/0962280213509720

תקציר 4

Estimation of Dwelling Values in Israeli Arab Sector

Yury Gubman and Larisa Fleishman

Central Bureau of Statistics

Abstract

The availability of register data on market value for the entire stock of residential properties is extremely important for economic and statistical analysis. While interpolation of property values data for Israeli Jewish sector has been recently accomplished for every dwelling recorded in the Dwelling Register, dwelling values for the entire population in Arab sector had not been estimated yet.

Arab housing market is characterized by limited rural to city movements, limited housing mobility, and rather limited size of a free housing market. Therefore, estimation of property value data for every dwelling in Arab sector is a methodological challenge.

In this paper we propose a method for the mass appraisal of residential properties in Israeli Arab sector. First, hedonic models were fitted to transaction prices acquired from the 2008-2011 Tax Authority records. These models were estimated for the sub-sample of localities with fairly developed housing market, in terms of total dwellings' stock, number of transactions, and advertising of homes for sale by owners. Rather limited housing market in other Arab localities justifies utilizing also subjective assessments of dwelling values culled from the 2008-2011 Household Expenditure Survey. We found that both transactions prices and owners' valuations are correlated with dwelling and locality features, as well as with location factors.

Following the presumption that the Household Expenditure Survey and the Tax Authority data are opposite biased, and no clear benchmark exists for dwelling values in the Arab sector, the prediction models were estimated using both data sources of property value information. Subsequently, two values for each dwelling were calculated. Then, the weighted average of the two estimated values was used as a final imputed value for the entire inventory of dwellings in the Arab sector. The weights for estimating final predictor of property value were calculated according to the number of transactions in a locality and its size, and calibrated compliant with the ranking of the localities by their average income.

Methodology developed in this study for estimating property value can be implemented annually. Apart from the statistical and administrative significance, estimation of property value for every dwelling in the Israeli Arab sector poses an important research challenge that may contribute to a more profound understanding of the distribution of physical residential properties in this sector.

Keywords: housing prices, value estimators, Arab sector

Modeling the Nonlinear Statistics of Cities in Israel

David Burg, Zinaida Ilatov

Golan Research Institute, University of Haifa, Katzrin

email: biomodel@research.haifa.ac.il

During the last two decades, urban research has witnessed an explosion of research tracking macroscopic social quantities in order to generate a "large picture" of social dynamics. Bettencourt and West et al. (2007, 2008, 2012) have developed a mathematical framework to describe the fundamental scale invariant dynamics of cities. They show that variables tracking infrastructure are a sublinear process and mature technologies are linear in nature. However, the main drivers for the success of cities are the apparent superlinear social-economic factors.

We calibrated their model with large datasets including multiple variables of cities and municipalities in Israel. Our results show similar patterns. For example, municipality expenditures, water usage and gas stations ($0.60 < exp < 0.95$) exhibit saturating dynamics. On the other hand, population density, energy, housing, solid waste increase exponentially, per capita ($1.05 < exp < 1.15$), as well as, fiscal variables like wages, the number of businesses and property taxes collected from them ($1.02 < exp < 1.5$). On the other hand, negative aspects of social structure, such as CO_2 emissions, traffic accidents, crime and deaths ($1.15 < exp < 1.42$), also increase exponentially as a function of the size of the city.

While the underlying fractal-like emergent self-organizing structures of cities in Israel may be similar, the dynamical rates differ significantly from the data found in the literature. This work has allowed a statistical description of the particular nonlinearities of Israeli urban structure. This has allowed us to derive a multi-variate model representing the characteristics of each city in relation to the expectation of the model. The importance is considerable for urbanization dynamics and strategic urban planning allowing more accurate projections which account for these dynamics.

בקרת איכות דו-רמתית על ציונים לפי משתנים דמוגרפיים

דביר קלפר, אבי אללוף, אליוט טורוול, כרמל אורן, מרינה פרונטון
מרכז ארצי לבחינות ולהערכה

מטרת המחקר הנוכחי היא לבדוק את הישימות של שלושה מודלים רבי-רמות לניבוי הציון בבחינה הפסיכומטרית על בסיס משתני רקע שונים, ואת האפשרות להסתמך על המודל האופטימאלי מתוכם בביצוע בקרת איכות של תהליך חישוב הציונים.

בהתאם לכך בוצע ניתוח היררכי לינארי בשתי רמות כדי לחקור את הקשר של נתוני הרקע לציון הכללי מעבר לנוסחים, כאשר הרמה הראשונה היא הנבחן הבודד, והרמה השנייה היא נוסח הבחינה.

שלושת המודלים שנבדקו הם:

1. ניתוח שונות עם אפקטים מקריים

2. רגרסיה של הממוצעים

3. מודל רגרסיה עם מקדמים מקריים

מודל 1 שימש כבסיס להשוואה עם שני המודלים האחרים. מודל 2 הצליח להסביר 69% מהשונות של הציון הממוצע ,לכן נסיק כי ברמת הנוסח יכולת ניבוי הממוצע היא טובה מאוד. מודל 3 הצליח להסביר 20% מהשונות של הציון של נבחן בודד, לכן אפשר להסיק כי ברמת הנבחן היכולת לחזות ציון לנבחן בודד היא חלשה יחסית. המודל האופטימאלי עם יכולת הניבוי הטובה יותר (מודל 2) נבדק על בָּסיס נתונים אחר ונמצא תקף.

מחקר זה הינו מחקר ראשון המשתמש בניתוח רב-רמתי כדי לחקור את הקשר בין משתני הרקע של נבחנים לבין הציון הכללי בבחינה הפסיכומטרית. תוצאות המחקר מעודדות ונותנות ביד החוקר כלי רב עוצמה לבקרת איכות של ציונים בבחינה זו ובמגוון רחב של מבחנים שבהם יש בידי החוקר נתונים דמוגרפיים של הנבחנים. בקרת האיכות יכולה להתבצע בדיעבד, על מבחנים שכבר הועברו, כדי לבקר את איכות חישוב הציונים והכיול, ובאופן שוטף, לאחר חישוב ציוני הבחינה וטרם דיווחם לנבחנים ולמוסדות המשתמשים, כדי לאתר בעיות בזמן אמת.

הערכת דיוקו של מדד מחירי המזון למרכיביו:
בחינה אמפירית באמצעות נתוני סורק

דורון סייג

דיוקו של מדד המחירים לצרכן, אינדיקטור סטטיסטי-כלכלי המחושב על סמך מדגם מחירים, תלוי באופן קריטי בהיקף ובאיכות המידע הגולמי המשמש לחישובו. לפיכך, ישנה חשיבות רבה בבחינת שיפור המידע הקיים ובבניית מסד נתונים מלא, מקיף ו"נקי" מטעויות ככל הניתן. לצד העלויות הגבוהות הכרוכות באיסוף מחירים בדרך המסורתית, קרי באמצעות איסוף של סוקר בחנויות (להלן: "נתוני סוקר"), יצרה הקידמה הטכנולוגית מקור נתונים חדש - מחירים אלקטרוניים מנקודות המכירה (להלן: "נתוני סורק"). מחקר זה בוחן על-פני 4 שנים את ההבדלים בין מדדי המחירים המתקבלים מנתוני סורק מול אלה המתקבלים מנתוני סוקר, ובמציאת הגורמים המרכזיים לשוני בין המדדים. חקירת הדיוק וההטיות נעשתה באמצעות בניית מדד ניסיוני מקביל, המתבסס על מסד נתונים בו שולבו על בסיס חודשי כ-200 אלף תצפיות מנתוני סורק. השיפור באיכות המידע הגולמי משתקף במכלול היבטים: גודל מדגם, מגוון המוצרים במדגם, כיסוי החנויות, דיוק בשיקוף של מבצעים והנחות. שיפורים אלו יחד עם שיפורים מתודולוגיים נוספים סיפקו בסיס איתן להנחת המחקר שהמדד הניסיוני משקף במדויק יותר את השינויים במחירים שחלו בשוק המזון בהשוואה למדד הרשמי. לצורך השוואת ההתאמה בין המדדים באופן מדעי ואובייקטיבי נעשה שימוש במדדי דמיון בין סדרות עתיות שפותחו למטרה זו. המחקר בוחן את הסדרות המרכיבות את סעיף המזון לכדי תיקוף סדרות מחירים העמידות למקור המידע החדש ומאתר סדרות מחירים בהן נמצאו פערים משמעותיים. בנוסף, מתבצע פירוק של סוגיות מתודולוגיות שיושמו במדד הניסיוני במטרה לאמוד את השפעתן על הממצאים. הסוגיות המתודולוגיות שנבחנו הינן: טעויות דגימה, הטיות הנובעות מתחלופת חנויות, אגריגציית מחירים על-פני זמן (מחיר שבועי ביחס למחיר תלת-שבועי) ושימוש במשקלות ברמת התצפית.