of the Symbolic Data Analysis will permit to solve all those critical aspects that until now have prevented a comprehensive description of the mechanisms of the static-structural evolution of the monument, and to simulate the possible "reactions" to environment changes of exceptional nature.

References

1 G. Bartoli, A. Chiarugi and V. Gusella, "Monitoring systems on historic buildings: the Brunelleschi Dome", *Journal of structural engineering,* 1997.

2 B. Bertaccini, "Santa Maria del Fiore Dome Behavior: Statistical Models for Monitoring Stability", *International Journal of Architectural Heritage: Conservation, Analysis, and Restoration,* 9:1, 25-37, 2015.

3 L. Billard and E. Diday, Symbolic data analysis: Conceptual statistics and data mining, Chichester: Wiley, 2006.

4 A. Cury, C. Crémona and E. Diday, "Application of symbolic data analysis for structural modification assessment", *Engineering Structures,* vol. 32, 762–775, 2010.

5 A. Cury and C. Crémona, "Assignment of structural behaviours in long term monitoring: application to a strengthened railway bridge", *Structural Health Monitoring,* vol. 11:4, 422–441, 2012.

6 J. Santos, C. Crémona, A. Orcesi, P. Silveira and L. Calado, "Static-based early-damage detection using symbolic data analysis and unsupervised learning methods," *Frontiers of Structural and Civil Engineering,* vol. 9:1, 1-16, 2015.

7 L. Billard, "Some Analyses of Interval Data", *Journal of Computing and Information Technology,* 16:4, 225-233, 2008.

8 J. Le-Rademacher and L. Billard, "Symbolic Covariance Principal Component Analysis and Visualization for Interval-Valued Data", *Journal of Computational and Graphic Statistics,* 21:2, 413-432, 2012.

9 K. Košmelj, J. Le-Rademecher and L. Billard, "Symbolic Covariance Matrix for Interval-valued Variables and its Application to Principal Component Analysis: a Case Study", *Metodološki zvezki,* 11:1, 1-20, 2014.

10 G. Alefeld and G. Mayer, "Interval analysis: theory and applications", *Journal of computational and applied mathematics,* 121, 421-464, 2000.

11 P. Bertrand and F. Goupil, "Descriptive statistics for symbolic data", in H-H. Boch and E. Diday (eds), *Analysis of Symbolic Data:Exploratory Methods for Extracting Statistical Information from Complex Data*, Berlin, Springer-Verlag, 106-124, 2000.

12 C. Blasi and G. Bartoli, "Il sistema di monitoraggio della cupola di Santa Maria del Fiore: problematiche relative al funzionamento degli strumenti e alla gestione dei dati", in *Monitoraggio delle strutture dell'ingegneria civile.*, Udine, CISM, 183-202, 1995.

13 D. Posenato, K. P. D. Inaudi and I. and Smith, "Methodologies for model-free data interpretation of civil engineering structures", *Computers and Structures,* vol. 88:7/8, 467-482, 2010.

14 W. Li, H. Yue and S. Valle-Cervantes, "Recursive PCA for adaptive process monitoring", *Journal of Process Control,* 10, 471-486, 2000.

15 X. Wang, U. Kruger and G. Irwin, "Process Monitorin Approach using Fast Moving WIndow PCA", *Industrial & Engineering Chemistry Research ,* vol. 44:5, 5691-5702, 2005.

16 B. Bertaccini, G. Biagi, A. Giusti and L. Grassini, "Symbolic data analysis approach for monitoring the stability of monuments", in M. Pratesi and C. Perna (eds.), *Proceedings of the 48th Scientific Meeting of the Italian Statistical Society*, 1-6, 2016.

A latent markov model approach for measuring national gender inequality

Modello latent markov per la misura delle disequità di genere nazionali

Gaia Bertarelli and Franca Crippa and Fulvia Mecatti

Abstract Gender inequality - both in space and time - is a latent trait, namely only indirectly measurable through a collection of observable variables and indicators purposively selected. Even if composite indicators are normally used by social-scientists, when measuring gender-gap they are known to have case-specific technical limitations. In this paper we propose an innovative approach based on a multivariate Latent Markov model (LMM) for the analysis of gender inequalities as measured by the aforementioned indicators.

Abstract *La Statistica di Genere si occupa di sviluppare metodologie atte a cogliere disparità e differenze nella situazione delle donne e degli uomini in tutti gli aspetti della vita. Negli ultimi anni le disponibilità di dati per l'analisi di genere è aumentata poiché sempre più paesi stanno adottando survey specifiche. Gli strumenti più comuni nella letteratura della statistica di genere sono gli indicatori compositi che tuttavia presentano note limitazioni metodologiche. Vogliamo proporre un approccio innovativo alla statistica di genere basato su un modello latent markov multivariato per le analisi delle disuguaglianze.*

Key words: Gender Statistics, Clustering, GID-Database OECD, latent variable.

1 Background and Introduction

Gender equality is a recognized goal of modern democracies and an objective for global civilization since the effects of policies and actions capable at reducing gen-

Gaia Bertarelli
University of Perugia, e-mail: gaia.bertarelli@unipg.it

Franca Crippa
University of Milano - Bicocca, e-mail: franca.crippa@unimib.it

Fulvia Mecatti
University of Milano - Bicocca, e-mail: fulvia.mecatti@unimib.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

der disparities would actually benefit the society as a whole, both women *and* men. The availability of good quality data for engendered statistical analysis at the national level has increased since the 90's. Gender statistics based on household surveys and administrative records are becoming widely available. Gender inequality is a latent trait, namely only indirectly measurable through a collection of observable variables and indicators purposively selected as micro-aspects contributing to the latent macro-dimension. This is one of the main reasons for the popular use of *composite* indicators as current gender statistics indicators, *i.e.* aggregations - usually linear combinations - of a collection of simple indicators each singled out for assessing a puctual micro-aspect of the latent gender dimension. Several world rankings, based upon national gender composite indicators, are periodically released by supranational agencies (see for instance [2] for a comparate review). Even if normally used by social-scientists, such gender-gap measures are known to have case-specific technical limitations [3], which often lead to internal inconsistency since the ranking of a single country can vary in relation to the indicator considered. Moreover, a significant amount of the literature criticizes the use of composite indicators on the ground of trivial marginalization and arbitrariness [4]. In this paper we propose an innovative approach to gender inequality measure based on a multivariate Latent Markov model (LMM).

2 Data

We focus on two inequality indexes, the Gender Inequality Index (GII) and the Global Gender Gap Index (GGGI). A main reason for selecting them is their recentness, whose the aforementioned technical issues ask for advanced knowledge. The GII, introduced by UNDP in 2010, measures gender inequalities in three aspects of human development: reproductive health, empowerment and economic status. The GGGI was first introduced by the World Economic Forum in 2006 as a framework for capturing the magnitude of gender-based disparities and for tracking their progress. Three basic concepts underlie the GGGI. First, the index focuses on measuring gaps rather than levels. Second, it captures gaps in outcome variables rather than in input variables. Third, it ranks countries according to gender gaps rather than women's empowerment. It measures four aspects: economic partecipation and opportunity, educational attainment, health and survival and political empowerment. Rankings based on these indicators are different from each other as well as not constant over time, as a consequence of different choices in both measurable variable selection and aggregation system. In this paper we consider a multivariate model of latent markov type, able to receive as input both indexes as well as a set of covariates. An improved gender inequality measure is expected as a result. A preliminary univariate analysis is conducted for the period 2010-2016 able to assess possibly measurement errors in GGGI and GII. After considering constitutional gender equity (see `http://constitutions.unwomen.org/en`) and social structure

as covariates in the latent model component, we introduce time use in the measurement part.

3 Model

LMMs (see [1] for a general review), are a class of statistical models for longitudinal data which assume the existence of a latent process which affects the distribution of the response variables. The existence of two processes is assumed: an unobservable finite-state first-order Markov chain $U_i^{(t)}$, $i = 1,\ldots,n$ and $t = 1,\ldots,T$ with state space $\{1,\ldots,m\}$ and an observed process $Y_i^{(t)}$, $i = 1,\ldots,n$ and $t = 1,\ldots,T$, where $Y_i^{(t)}$ denotes the response variables for area i at time t and similary for $U_i^{(t)}$. We assume that the distribution of $Y_i^{(t)}$ depends only on $U_i^{(t)}$: the latent process fully explains the observable behaviour of an item together with possibly available covariates. Therefore it is important to distinguish between two components: the measurement model, which concerns the conditional distribution of the response variables given the latent process, and the latent model, which concerns the distribution of this latent process.

The unknown vector of parameters ϕ in a LMM includes both the parameters of the Markov chain ϕ_{lat} and the vector of parameters of the state-dependent distribution ϕ_{obs}. The *measurement model* involves ϕ_{obs} and it can be written as $Y_i^{(t)}|U_i^{(t)} \sim f(y,u,\phi_{obs})$. The *latent model* includes the parameters ϕ_{lat} of the Markov chain which are the elements of the transition probability matrix $\Pi = \{\pi_{u|\overline{u}}\}$, with $u,\overline{u} = 1,\ldots,m$; where $\pi_{u|\overline{u}} = P(U_i^{(t)} = u|U_i^{(t-1)} = \overline{u})$ is the probability that area i visits state u at time t given that at time $t-1$ it was in state $\overline{u}$, and the vector of initial probabilities $\pi = (\pi_1,\ldots,\pi_u,\ldots,\pi_m)'$ where $\pi_u = P(U_i^{(1)} = u)$ is the probability of being in state u at the initial time for $u = 1,\ldots,m$. In this work we consider homogeneous LMMs.

LMMs can assess the presence of measurement errors or account for unobserved heterogeneity between areas in the analysis including covariates in the measurement model which do not completely explain the heterogeneity in the response variables. In LMMs the effect of the unobservable variable has its own dynamics. Moreover, a latent clustering of the population of interest can be pointed out. Our proposal is based on adapting the LMM to the gender statistics framework by interpreting national gender gap as the latent status of interest and using the distributions of the GGGI and GII as response variables. This metodology is derived by integrating into the same LMM both the selected composite indicators and a set of available observable covariates of any and possibly mixed nature. Our metodology organizes countries in ordinal clusters representing of the severity of gap. The classification is produced taking into account the values of the considered covariates and this overcomes the so called "world-at-two-speed" effect, i.e gender inequalities due to the denial of basic human rights (under-developing or in transition coutries) or due to uneven opportunities between men and women (developed countries with gender

equality stated by law) ([2])) which is evident especially in the GII's distribution. However, looking at the temporal distributions, it seems that this gap goes to dwindle with time. Because of this, a longitudinal analysis is appropriated. We conduct a two-step analysis. At the beginning we apply a LMM with only spacial and gender constitutional equality covariates on the latent model in order to identify clusters of countries actually comparable under the "two-speed" effect mentioned above. Then we apply a LMM within each cluster considering social and economic covariates in the measurement model to detect main differences and variability within the same group.

4 Expected Results

We propose to integrate into the same LMM both the selected composite indicators and a set of available observable covariates of any and possibly mixed nature, categorical, ordinal and quantitative, fully exploiting the multidimensional latent nature of gender imbalance. The model would provide an organization of the countries in a (optimal) number of ordered cluster. The classification is produced taking into account the values of the considered covariates and this overcomes the so called "world-at-two-speed" effect which is evident especially in the GII's distribution. Moreover the proposed methodology deals with the forecasting of the future response and the path prediction.

References

[1] Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. *Latent Markov models for longitudinal data*. CRC Press, 2012.

[2] Fulvia Mecatti, Franca Crippa, and Patrizia Farina. A special gen (d) re of statistics: roots, development and methodological prospects of gender statistics. *International Statistical Review*, 80(3):452–467, 2012.

[3] Iñaki Permanyer. The measurement of multidimensional gender inequality: continuing the debate. *Social Indicators Research*, 95(2):181–198, 2010.

[4] Martin Ravallion. On multidimensional indices of poverty. *The Journal of Economic Inequality*, 9(2):235–248, 2011.

Eurostat's methodological network: Skills mapping for a collaborative statistical office

Agne Bikauskaite and Dario Buono

Abstract Collaboration, interaction and exchange of knowledge among staff are important components for development and enriching of scientific intelligence within a statistical office. Eurostat's methodological network has been built as a skills mapping tool aiming identify in-house competencies for innovation and affordability of diffusion of knowledge, promotion and modernisation of collaboration on methodological issues, and processes within statistical office. We mainly focus on staff's knowledge and working and academic experience in methodological areas, domains and tools on statistics and econometrics. Quantitative network analysis metrics are used to measure the strengths of existing methodological competencies within Eurostat, to identify groups of people for collaboration in providing results on specific tasks, or characterise areas that are not fully integrated into methodological network. By combining network visualisation and quantitative analysis, we able easily assess competency level for each dimension of interest. Network analysis helps us in making decisions related to improvement of staff communication and collaboration, by building mechanisms for information flows, filling competency gaps. Data represented as mathematical graph makes readily visible general view, absorbs its structure, permits us to focus on persons, competencies and relations between them. Modernisation of ways of working leads to a more cost effective use of existing resources.

Key words: complex network, data analysis, network visualization, bipartite graphs, network projection, ego network, network analysis

1 Introduction

Collaboration, interaction and exchange of knowledge among staff are important components for development and enriching of scientific intelligence within a statistical office, especially when this exchange happens across areas of interest by both interacting sides. Methodological network has been built as a skills mapping tool aiming identify in-house competencies for innovation and affordability of

[1] Agne Bikauskaite; email: agne.bikauskaite@ext.ec.europa.eu
Dario Buono, Eurostat; email: dario.buono@ec.europa.eu

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

diffusion of knowledge and information, and promotion and modernisation of collaboration on methodological issues and processes within statistical office. We mainly focus on staff's knowledge and working and academic experience in methodological areas, domains and tools on statistics and econometrics. This paper provides a set of mathematical network analysis measures from basic ones as size and degree to more complex as clustering coefficient and their correlation with degree that evaluates and makes better understandable the methodological knowledge network structure.

2 Methods

Quantitative network metrics are used to measure the strengths of existing methodological competencies within statistical office, to identify groups of people for collaboration in providing results on specific tasks, or characterise areas that are not fully integrated into methodological network. Network analysis helps us in making decisions related to improvement of staff communication and collaboration, by building mechanisms for information flows, filling competency gaps. By combining network visualisation and quantitative analysis, we can easily assess competency level for each dimension of interest.

2.1 Bipartite graph

Network data consists of a set of elements with relations on those elements and it may be represented as a graph. Our research subjects, individuals, form links which characterise their competencies in statistics and econometrics. Formally we have a graph $G = (V, E)$, where G is a relational structure consisting of set of vertices V and set of edges E [2]. We say that a graph is bipartite when the vertex set V is divided into two finite, disjoint $V_1 \bigcap V_2 = \varnothing$ sets [4]. When V_1 composed of the first mode vertices and V_2 of the second mode vertices, we have the bipartite graph $G = (V_1, V_2, E)$ where ties map the elements of different modes only.

2.2 Network analysis

In order to understand organisational methodological network and its structure network analysis statistical models have been employed. Data arranged as person by skill matrix A of size $n_{V_1} \times n_{V_2}$, where the rows correspond to methodological

network members interested in collaboration and knowledge exchange and the columns to the dimensions of competencies:

$$A_{ij} = \begin{cases} 1, \text{if person } i \text{ has a link to methodological skill } j; \\ 0, \text{otherwise.} \end{cases}$$

The two most basic parameters of the graph are the number of vertices $n = n_{V_1} + n_{V_2}$, where $n_{V_1} = |V_1|$ and $n_{V_2} = |V_2|$, and the number of edges $m = |E|$. [3]

Degree of the vertex helps to identify the best known competencies, and to diagnose critical areas within the methodological network. The average degree of sets of vertices V_1 corresponding to survey respondents and V_2 characterising listed methodological competencies are commonly used summarizing how well connected the network is, and is defined as proportions of number of links the network and number of nodes [1]

$$k_{V_k} = \frac{m}{n_{V_k}}, \text{where } k = 1,2.$$

While the average degree of overall network is obtained from the total numbers of nodes and edges by following equation [1]

$$k = \frac{2m}{n_{V_1} + n_{V_2}}.$$

The density δ of the bipartite graph G measures average ratio of the actual degree of the nodes in the network and the maximum possible degree, which corresponds to the number of nodes in the set of different mode nodes

$$\delta(G) = \frac{m}{n_{V_1} n_{V_2}}.$$

This index is equal to 1 for the fully connected network (i.e. G has one component) and takes value of 0 when network is fully disconnected (i.e. G is composed entirely of isolates).

The clustering coefficient which concerns link correlation gives an idea of how compact is the network. The clustering coefficient of a node i is the proportion of links between the nodes within its neighbourhood divided by the number of edges that could possibly exist between the nodes [4]

$$cc_{ijl} = \frac{q_{ijl}}{(k_j - \eta_{ijl}) + (k_l - \eta_{ijl}) + q_{ijl}}$$

here j and l are a pair of neighbours of node i, q_{ijl} is the number of squares which include these three nodes, and $\eta_{ijl} = 1 + q_{ijl} + \theta_{jl}$ with $\theta_{jl} = 1$ if i neighbours j and l are connected with each other and 0 otherwise.

Existing correlation of links allows us to sustain collaboration between methodological network members, while otherwise would not be able to function. If persons i and k form links to common competencies j and l, then efficient cooperation between them is more likely possible.

3 Results

The methodological knowledge network of this study case is simple, undirected, unweighted, static, and structured as bipartite graph, which consist of 117 vertices connected by 595 edges. The competencies degree of staff participated in the survey ranges from 3 to 11 which a mean of 8.88. While the degree of competencies nodes ranges from 0 to 39, with a mean of 10.2, what indicates, that each competence from the list has been indicated as well known by 10 respondents on average.

Degree sequence of competencies in statistics and econometrics indicates that most of methodological network members are familiar to Data Analysis and Time Series, highly competent in Social Statistics and National Accounts, and experienced in R and SAS statistical analysis software. While the biggest gap within methodological network observed of experts on Micro-data access and Statistical confidentiality, knowledgeable in Transport and Energy statistics, and capable to work with Hadoop tool. Other competencies from defined list are more or less covered and known by methodological network members.

The standard density measure gives a value 0.17, which shows a fairly sparse network with presence of 17 per cent of the possible links for average node. However, in this particular case the standard denominator is clearly not appropriate defining methodological network members' competencies. Due to restriction of choice of maximum 11 dimensions out of 50 possible, it cannot be interpreted as actual possible density. Using modified denominator, network obtain density of 0.79, which indicates high competency level of methodological network members.

In network studied, the clustering coefficient of competencies vertices set is not so high, above 20 per cent. The moderate correlation between clustering coefficient and degree is detected.

Data represented as mathematical graph makes readily visible general view, absorbs its structure, and permits us to focus on persons, competencies and relations between them. We distinguish the two node sets by colours, so that nodes of the same type have the same colour. The vertices of staff willing collaborate are coloured in green, blue, and red depending on the type of interest in involvement, while the set of yellow vertices corresponds to 28 methodological areas, 12 statistical domains and 10 tools. The size of the label and vertex is proportional to its degree.

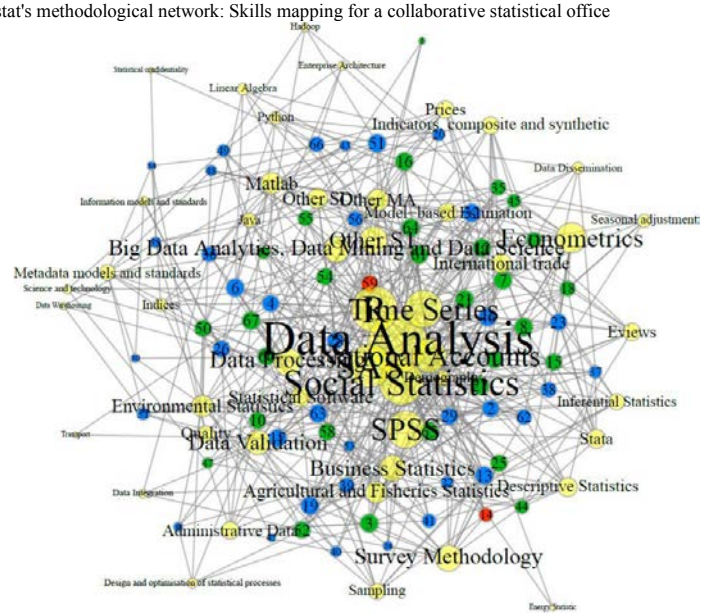

Figure 1: Organisational methodological knowledge network

In order to simplify visualisation, for deeper analysis of existing knowledge features and easier identification of clusters of correlated areas, methodological network has been divided into sub-networks by different breakdowns. Projections into one mode networks to grasp weighted relations between the same set of vertices had been made available as well by multiplying matrix A and its transpose A'. Analysing sub-networks we notice the tendency of increase of the density when average degree decreases. Overlapping of the structure of the nodes is very small, what points that there is large community of the methodological network members with knowledge and skills in different variation of areas.

4 Conclusions and discussion

In this study we map and evaluate existing methodological skills within the statistical office applying network analysis techniques. Networks as analytical and visualisation tools provide a number of useful outcomes. By detecting and then mapping methodological skills within organisation we are able to understand, spread, monitor and maintain existing skills, to develop tools for better knowledge accessibility and modernise information diffusion ways.

Obtained results provide quantitative evidence that methodological network members are qualified in different areas, given measures ensure possibility of well collaboration performance within the statistical office. Network is highly connected, significant gap of competencies is detected only in one methodological area from the defined competencies list of interest.

We can outline the importance of detecting and monitoring existing knowledge and skills within modern statistical office. Two employees could affect each other only if they know about each other and that common competencies are available between them, as efficient communication and collaboration within the organisation is possible only when we know with whom we could potentially contact. As well modernisation of the statistical office's ways of working leads to a more cost effective use of existing resources. Network is a key source in promoting and supporting of knowledge diffusion and expanding, enriching professional and personal skills and filling knowledge gaps within statistical office.

References

1. S. P. Borgatti, M. G. Everett (1997). Network analysis of 2-mode data. Social networks, 19, 243-269
2. C. T. Butts (2008). Social network analysis: A methodological introduction. Asian Journal of Social Psychology, 11, 13-41
3. R. A. Hanneman, M. Riddle (2005). Introduction to social network methods. http://faculty.ucr.edu/~hanneman/nettext/index.html
4. M. Latapy, C. Magnien, N. Del Vecchio (2008). Basic notions for the analysis of large two-mode networks. Social Networks, 30, 31-48