

A Data-driven Approach for the European Gender Equality Index

Lorenzo Giammei¹ , Flaminia Musella² , Fulvia Mecatti³ , Paola Vicard⁴ 

¹Department of Social Sciences and Humanities, National Research Council, Italy, ²Education Department, Roma Tre University, Italy, ³ Department of Sociology and Social Research, Milano-Bicocca University, Italy, ⁴Economic Department, Roma Tre University, Italy.

How to cite: Lorenzo, G.; Flaminia, M.; Fulvia, M.; Paola, V. 2025. A Data-driven Approach for the European Gender Equality Index. In: 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025). Rome, 2-4 July 2025. <https://doi.org/10.4995/CARMA2025.2025.20469>

Abstract

Gender equality is a fundamental human right and an objective in the United Nations Agenda 2030 for sustainable development. Assessing the gender gap usually relies on composite indicators: tailored statistical tools that are effective in summarizing a set of indices in a single number. However, the availability of regional microdata can open the way to statistical learning tools that leverage the information contained in big structured datasets to allow deeper analyses. In this work we employ an Object-Oriented Bayesian Network to measure the gender gap on Italian province level data. The model is consistent with the European Gender Equality Index, while enabling the investigation of multivariate interactions and the simulation of scenarios. The proposed approach shows how statistical learning can enrich traditional composite indicator analysis and shed light on the determinants of gender inequality.

Keywords: *Statistical learning; Gender Equality Index; Object-Oriented Bayesian Network; Composite indicator; Gender gap.*

1. Introduction

The European Institute for Gender Equality (EIGE) works to promote gender equality within the EU and beyond by serving as a European knowledge center on gender equality, supporting gender mainstreaming in EU and Member State policies, and combating the gender gap. EIGE monitors and compares European countries by the Gender Equality Index (GEI), a composite indicator that ranks countries with respect to their levels of gender inequality, fostering the reduction of the gender gap in the European Union (EU). The computation of the index is based on a set of indicators, also called ingredients, purposely chosen to reflect gender disparities across a selection of dimensions. The ingredients are then adjusted and aggregated to compute first subdomain and then domain scores. Each domain represents a dimension of gender inequality, namely Work, Money, Knowledge, Time, Power, and Health. The obtained domain scores are then combined linearly to produce the final GEI score, which ranges from 1 (representing maximum inequality) to 100 (indicating perfect equality).

Figure 1 shows a map of European countries colored based on their 2023 GEI value; the darker the color, the higher the index. Italy's GEI score is 68.20, placing it 13th in Europe.

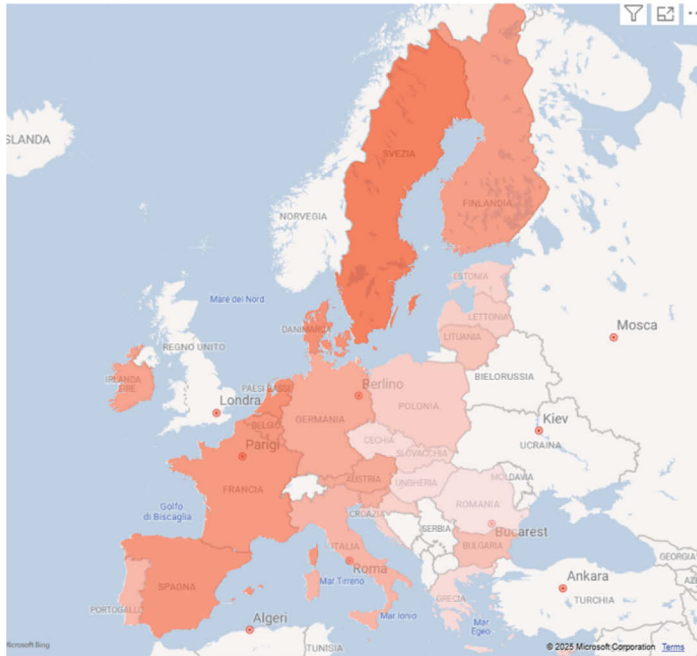


Figure 1. Map of European countries, darker colors correspond to higher GEI scores. Source: Elaboration on EIGE data (2023).

Although this score is below the European average of 70, Italy has seen the most substantial improvements among EU countries, with a percentage increase of 28% over the past decade.

The computation of a GEI score requires country level data, however if gender gap indicators are available at regional level, the index can be also computed for smaller geographical areas. For example, Fachelli and López-Roldán (2023) discuss the construction of a regional GEI for Italy, Spain, and France, analyzing gender equality across 62 regions. In the paper they highlight the importance of regional disparities and the need for tailored local gender policies. In a similar fashion, Di Bella et al. (2020) compute the GEI of Italian regions. Their work addresses the need for sub-national measurements of gender equality to guide regional policy priorities and efforts.

The strength of gender composite indicators can be effectively leveraged through micro-regional applications, especially in countries marked by substantial heterogeneity, as is the case in Italy. However, the computation still provides a single number for each region, ignoring relations among indicators, i.e. without addressing multivariate interactions influencing the gender disparities. In this work, we propose a different approach to analyzing regional gender data, building on Mecatti et al. (2022), Musella et al. (2023), and Giammei et al. (2023). Rather than considering each region separately, we aggregate local data to form a larger structured dataset that can be used to train a statistical learning model. We argue that the use of artificial intelligence in data analysis, particularly statistical learning, could play a crucial role in addressing gender disparities and guiding gender-specific policies. To this end, we employ a statistical learning approach, specifically Object-Oriented Bayesian Network (OOBN), to model the GEI score, shed light on its generating process and conduct scenario simulations.

2. Object-Oriented Bayesian Networks

In this section we introduce the methodological foundations of OOBNs and show how they can be employed to model and empower a composite indicator.

First, Bayesian networks (BNs) must be defined. A BN is a multivariate statistical model encoding variable relationships through a Directed Acyclic Graph (DAG) and a joint probability distribution. A graph is a mathematical object composed by a set of nodes and a set of edges. Each node of the DAG is associated with a random variable and the edges constitute the relational structure between variables. In particular, the absence of an arc between two nodes indicates conditional independence between the corresponding variables. The DAG is acyclic in the sense that it is not possible, starting from a node, to end up in the same node following arrows directions. Each node is associated with the conditional probability distribution for the corresponding variable, say x_i given its parents, i.e. the set of nodes with an outgoing edge pointing towards x_i . The joint probability distribution can be factorised according to the DAG as follows:

$$P(x_1, \dots, x_p) = \prod_i P(x_i | pa_i), \quad (1)$$

where pa_i indicates the parent set of variable x_i , $i=1,..p$.

When subject matter knowledge is partial or missing, as is often the case in practice, the BN model can be learned from data employing structural learning algorithms (Kitson et al., 2023). These algorithms are data driven and, under specific conditions, generate either a DAG or a set of equivalent DAGs as output. When available, prior knowledge about the subject matter can be incorporated into the structural learning process. Once learned the graph's structure, the parameters that define the conditional probability distributions associated with the BN are estimated using the EM algorithm, leading to maximum likelihood estimates.

An Object-Oriented Bayesian Network (OOBN) is a hierarchical network made up of modular objects, or sub-networks. In addition to the typical nodes found in BNs, referred to as chance nodes, these networks also include instance nodes. An instance node is a modular object, which is a BN itself. Instance nodes exchange information with the other nodes/objects of the network through interface nodes, called input and output nodes. Notably, since OOBNs are constructed as a hierarchy of BNs, all of their statistical properties are preserved. Specifically, inference within the network is made more efficient by leveraging the conditional independence between chance nodes within the same instance and the broader OOBN, given the interface nodes.

OOBNs capability of modelling variable relationships in a modular, hierarchical framework, makes them an ideal statistical tool to faithfully represent and model a composite index architecture. At the same time, they simultaneously enable a nuanced exploration of the multivariate interactions underlying the index. In contrast to composite indicators—which consolidate several measures into a single score—an OOBN significantly enhance both explainability and interpretability. Its sophisticated inference engine reveals how modifications in one variable influence others by explicitly modeling the interrelationships among variables. Additionally, the inherent modularity of OOBNs permits the incorporation of extra layers beyond the composite indicator, offering enhanced analytical flexibility. Ultimately, these features can help to inform policies to reduce gender gaps by pinpointing priority areas for intervention.

3. An OOBN to model the Italian GEI from province level data

Italian province level gender data of 2019 have been employed to learn an OOBN that mimics the architecture of the GEI. It was not possible to cover all the original GEI domains due to lack of official statistics at this finer geographical level. However, with respect to the original version of the composite indicator, the model was enriched with some extra variables providing supplementary information on the socio-economic context at province level. Leveraging this dataset, an OOBN was learned using the Necessary Path Condition (NPC) algorithm which

combines automatic learning with domain-specific expertise and allows users to choose the most appropriate model among equivalent alternatives. The obtained model is displayed in Figure 2.

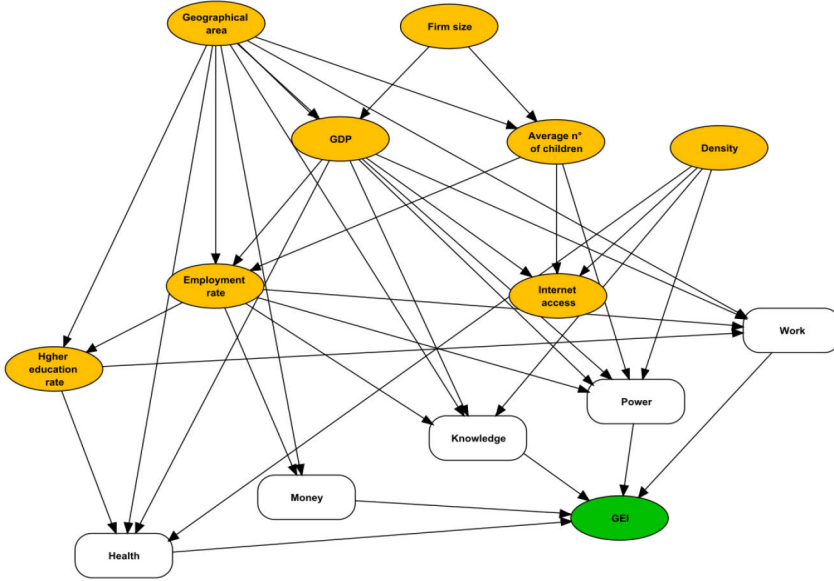


Figure 2. Object-Oriented Bayesian Network of Italian GEI at Province level.

Within the obtained model, extra variables are represented as (oval shaped) orange chance nodes and are connected to instance nodes, depicted as white smooth rectangles. Each BN instance represents a domain, therefore it is a BN itself including chance nodes, namely gender gap indicators, and interface nodes that connect the instance with the rest of the OOBN. Instance nodes have multiple input nodes and a single output node corresponding to the (dark green colored) deterministic target node GEI. The extra variables assigned as input of the instance nodes were selected based on a data-driven model, whose goodness-of-fit was assessed through prediction ability, measured with the ROC curve. This GEI target node's empirical distribution is defined by the weighted arithmetic mean of domain scores, with weights reflecting EIGE experts' opinion (Barbieri et al., 2017). Note that Time domain is omitted due to insufficient data at province level.

Beyond computing the GEI, the OOBN inferential engine can be exploited for scenario simulation, also known as what-if analysis allowing to assess how changes in the level of one or more variables affect the probability distribution of the other variables in the model. Therefore, the OOBN tool can support the decision maker in carrying out ex-ante evaluation, and for identifying the priority domain to reduce gender disparities, revealing viable strategies to achieve a certain level of GEI. In Figure 3 we see the probability distribution of the GEI,

together with the probability distributions of Financial resources equality, education participation equality and employment rate equality (from now on node names are reported in *italic*), that are three ingredient variables within Money, Knowledge and Work domains respectively. The GEI node distribution shows that its lowest level, describing a score between 0 and 50, is associated with a probability of almost 0.7. Such a high probability value reflects the information contained in the data and sheds light on the negative performance of Italian provinces in terms of gender gap.

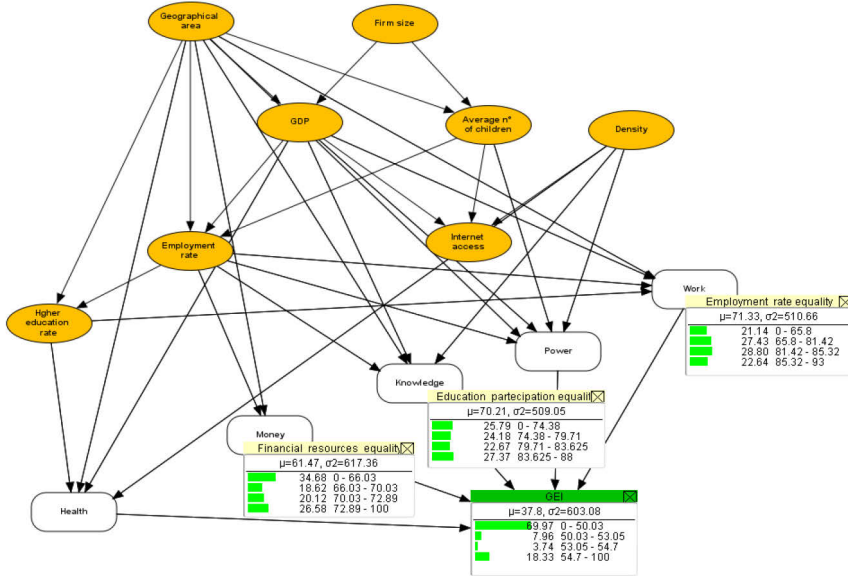


Figure 3. Object-Oriented Bayesian Network of Italian GEI – Baseline scenario.

As an example of scenario simulation, suppose we are interested in investigating what would happen if Education participation equality would rise to its maximum level. In other words, we assume to observe almost equal rates of female and male university enrollment rates. Computationally, this means that the probability of observing the highest level of education participation equality is set to 1. This evidence is thus inserted in the model and it is propagated throughout the OBN. Inserting the mentioned evidence is pictured in Figure 4. as a red bar appearing in the monitor probability window of the related node. Finally, all the probability distributions of the remaining variables are updated. Simulation results are shown in Figure 4.

Looking at GEI node first, we can remark that the probability associated with the lowest GEI score class (0-50) shrank, in favor of an increase in the probabilities of the other levels. In particular, the probability of observing a GEI between 54.7 and 100 rises from 0.18 to 0.25 (38.9% increase), denoting a relevant improvement in overall gender equality.

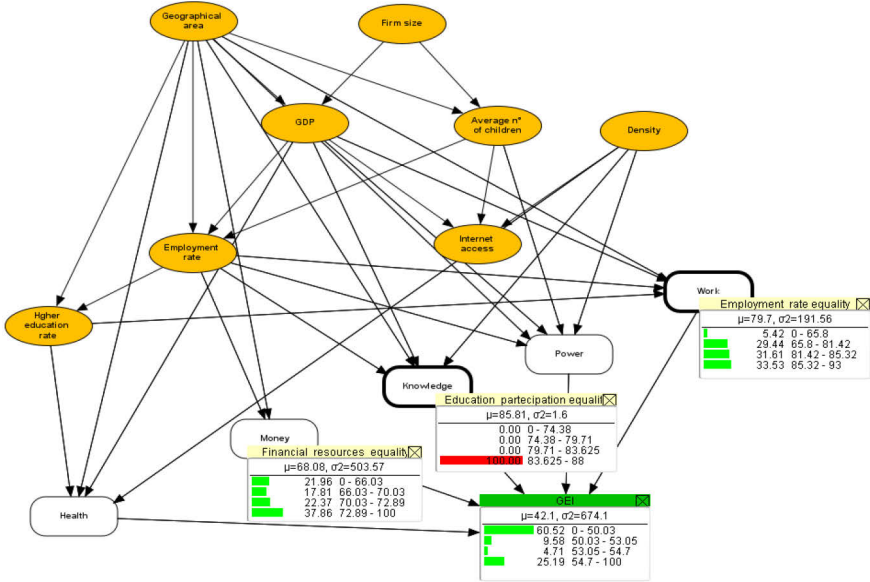


Figure 4. Object-Oriented Bayesian Network of Italian GEI – Simulated scenario.

It is important to underline that the described improvement in the GEI score is not only due to the mere increase in Education participation equality but, instead, results from a systemic effect of the provided evidence across the entire OOBN. We can in fact notice how, when we assume to observe an equal university enrollment rate between men and women, as we would expect, also Employment rate equality and Financial resources equality benefit from this shift. The improvement in the GEI distribution, is thus also driven and strengthened by the chain of indirect effects induced by the initial evidence.

4. Open Issues

In this work, we have shown how sub-national gender data can be aggregated to learn a statistical model that embeds the architecture of a composite indicator, while shedding light on the interactions that contribute to its computation. Transitioning from a composite index to a statistical learning model implied a fundamental shift in how indicator data is typically handled. Generally, computing a composite indicator, such as the GEI, requires only a single data point for each dimension covered by the index. In contrast, a statistical learning model like an OOBN benefits from a large volume of structured data points, from which information can be extracted and encoded into an advanced multivariate system. In this context, the availability of gender data at a local scale provided a foundation for reimagining country-level gender equality measurement from a big data perspective. The proposed OOBN offers additional and complementary insights compared to composite indicators. While the latter serve as

straightforward tools that prioritize synthesis and clarity, OOBNs reveal the relational structure between indicator dimensions, enabling informed policy guidance. Moreover, OOBN scenario simulation allows to go beyond composite indicator static computation, by investigating how all the interactions between variables contribute to shaping a certain GEI level, given a specific piece of evidence. The approach is highly generalizable and, when data is available, it could be extended to any other composite indicator presenting a hierarchical relational structure. Other gender composite indicators could thus be modeled through OOBNs and the obtained results could be compared across indices and methods.

The model could be further refined by incorporating a larger training sample and a broader set of indicators. However, in the case of Italian gender data, province-level granularity was the finest available. Future developments could also explore the inclusion of additional extra variables to construct a more comprehensive picture of gender equality in Italy.

References

- Barbieri, D., Karu, M., Lestón, I. R., Mollard, B., & Reingardé, J. (2017). *Gender Equality Index 2017—Methodological Report. Tech. rep.* Luxembourg: Publications Office of the European Union
- Di Bella, E., Leporatti, L., Gandullia, L., & Maggino, F. (2020). Proposing a regional gender equality index (R-GEI) with an application to Italy. *Regional Studies*, 55(5), 962–973.
- Fachelli, S., & López-Roldán, P. (2023). Comparative Analysis of Regional Gender Disparities. In E. di Bella, S. Fachelli, P. López-Roldán & C. Suter (Eds.), *Measuring Gender Equality* (pp. 135–166). Cham, Switzerland: Springer.
- Giammei, L., Musella, F., Mecatti, F., & Vicard, P. (2023). Building improved gender equality composite indicators by object-oriented Bayesian networks. In P. Coretto, G. Giordano, M. La Rocca, M. L. Parrella, & C. Rampichini (Eds.), *Book of abstracts and short papers - 14th Scientific Meeting of the Classification and Data Analysis Group Proceedings* (pp. 494–497). Italy: Pearson.
- Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., & Chobtham, K. (2023). A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, 56(8), 8721–8814.
- Mecatti, F., Vicard, P., Musella, F., & Giammei, L. (2022). Bayesian networks versus gender bias. *Significance*, 19(5), 16–20
- Musella, F., Giammei, L., Romio, S., Vicard, P., & Mecatti, F. (2023). Bayesian networks as a territorial gender impact assessment tool. In F. M. Chelli, M. Ciommi, S. Ingrassia, F. Mariani, & M. C. Recchioni (Eds.), *Intermediate Scientific Meeting of the Italian Statistical Society Book of short papers* (pp. 1084–1087). Italy: Pearson.