# 14TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS


PROCEEDINGS OF geoENV2022

**Andrea Zanini & Marco D'Oria**, Editors

UNIVERSITÀ DI PARMA

# UNPACKING OCCUPATIONAL HEALTH DATA IN THE TERTIARY SECTOR. FROM SPATIAL CLUSTERING TO BAYESIAN DECISION MAKING

María Pazo (1)* - Carlos Boente (2) - Teresa Albuquerque (3, 4, 5) - Natália Roque (3, 4) - Saki Gerassis (1) - Javier Taboada (1)

*GESSMin Research Group, Department of Natural Resources and Environmental Engineering, University of Vigo, Vigo, Spain (1) - CIQSO-Center for Research in Sustainable Chemistry, Associate Unit CSIC-University of Huelva "Atmospheric Pollution", Huelva, Spain (2) - Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal (3) - Centro de Estudos de Recursos Naturais, Ambiente e Sociedade (CERNAS), Instituto Politécnico de Castelo Branco, Portugal (4) – ICT, Universidade de Évora, Évora, Portugal (5)*
** Corresponding author: sakis@uvigo.es*

## Abstract

The health status of the service sector workforce is a great unknown for medical geography. Despite the advances carried out by spatial epidemiology to predict spatial patterns of disease incidence, there are important challenges unsolved. In particular, the main issue resides in the ability to effectively simplify and visually represent the problem domain, given the need to cover very different service activities and, at the same time, consider the impact of numerous emerging risk factors such as those stemming from bioclimatic and socioeconomic variables. This article proposes a new approach that allows to consider, simplify, prioritise and visualise multiple occupational health risk factors giving rise to not healthy workers. For that, it is used a twofold approach based on an innovative synergy between Bayesian machine learning and geostatistics, to analyse up to 74.401 occupational health surveillance tests gathered between 2012-2016 in Spain. This solution allows to extract relevant patterns over those risk factors that cannot be further discriminated in the Bayesian network, such as *spine* or *limbs observations*, depicting distribution maps of key differentiating variables computed by an ordinary kriging approach.

Keywords: Health data; Information theory; Ordinary kriging; Target analysis.

## 1. Introduction

The service sector, generally known as the tertiary sector of the economy, consists of the provision of services to other businesses, including end consumers. Services generate approximately 70% of the European Union's Gross domestic product (GDP) and employment (Eurostat, 2022). Some of the most common areas of the service sector are tourism (e.g., accommodation, travel agents), catering, education, real state, transport, and financial-related services. The variety of possible activities within this sector makes extremely complex the estimation of the health status of their workforce.

This aspect has been accentuated by the impact of the COVID-19 pandemic (Chang et al., 2021). Overall, men's and women's work tasks are in many cases considerably different, triggering occupational health risks for each gender. Despite the advances carried out by spatial epidemiology methods to predict spatial patterns of disease incidence, the abundance and accuracy of occupational health risk maps are still very limited (Gerassis et al., 2021). This is due in part to the multiple variables to represent without a clear approach to simplify the problem domain, and even more, to find out those differentiating variables. To this situation, it must be added

the already undeniable impact on the health of climate change (Orlov et al., 2020). The rise in temperatures is expected to open the door for an increasing number of pathologies whose effects can worse at work.

In this respect, given the outcomes of numerous investigations related to the effect that the climate change has on morbidity, reduced productivity of people, and increased sick leaves (Ebi et al., 2021; Wondmagegn et al., 2021) is necessary to bring a new perspective that addresses these challenges. For that, this study aims to develop an innovative approach, based on a methodological decision-to-visualization process that bears into consideration the impact of bioclimatic and socioeconomic variables as any other medical variable as part of the decision making process of a worker health status and associated occupational risks.

In practice, this research aims to improve the projections of occupational health risk factors and the characterization of the health status, which is the target node of the model, exploiting the combination of Bayesian machine learning and spatial techniques. In that manner, the added value is the possibility to identify and characterise those variables that may have a differentiating impact that apparently is not meaningful from a mathematical point of view. All in all, the results of this research work are expected to be one more contribution towards the medical services of the future, where the patient health status will not be any more subject to only a series of traditional medical tests and underlying medical conditions (Awotunde et al., 2021).

## 2. Material and Methods

### 2.1. Data characterization

A total of 74,401 occupational health surveillance tests gathered from workers belonging to the service sector in the period between 2012-2016 throughout the Spain territory were used as a medical data source for this study. More specifically, the workers for this research database carried out activities related to administrative and auxiliary services (31,894), financial and insurance services (12,958), education (13,938), and hostelry (15,611). Each clinical examination was undertaken according to the Spanish occupational health legislation (Ley 31/1995). Relevant occupational health organizations and hospital services conducted the medical tests gathering major information about the state of workers' health defining the main health risk factors causing pathologies, including the main physical conditions and health habits.

This study goes a step beyond traditional occupational health surveillance analyses, adding to the medical record of each worker a cross prediction with climatic and socioeconomic factors as an instrument to better characterize and predict those factors disrupting the health status. For that, *Maximum Temperature (BIO5)* or *Annual Rainfall (BIO12)*, and *Unemployment Rate or GDP* are examples of the bioclimatic and socioeconomic variables used respectively. Procedurally, this research is conducted in four levels. First, from the 37 initial variables considered, a total of 26 were finally taken for modeling purposes after reducing the problem dimension (Level 1). In the second level of analysis, these reduced variables were used to characterize the four main groups of service activities (Level 2). Later, for each activity group, the health status acts as a target node for which the relevant patterns are ascertained (Level 3). Figure 1 provides a scheme of this methodological process applied. These three levels presented correspond to the development of a Bayesian methodology that is complemented with a geostatistical analysis (Level 4) for those parts of the network where further clarity is needed.

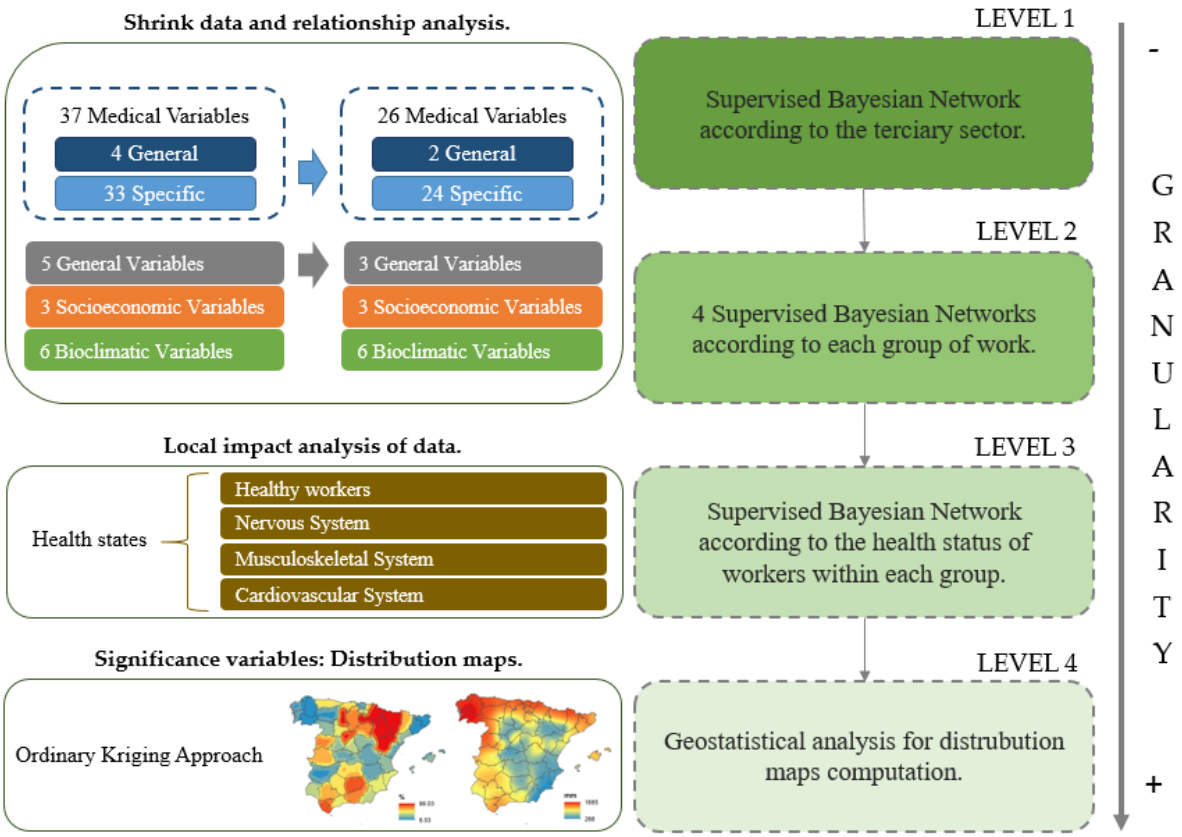


Figure 1 – The implemented methodological process with four levels of analysis.

As anticipated, for those occupational health variables that cannot be further discriminated in the network, a higher level of granularity in the analysis is provided by carrying out a geostatistical ordinary kriging approach as an effective solution to extract relevant patterns and produce reliable health risk maps. Ordinary kriging is the most widespread method of kriging. It serves to estimate a value at a point of a region for which a variogram is known, using neighboring data to the estimation of an unknown location (Goovaerts, 1997). This approach allows focusing the spatial representation on those variables with a more differentiating nature across the four different groups of service activities under analysis. Spatial interpolations were carried out by means of Geostatistical Wizard module in ArcGIS v-10.2.2. Semivariograms were manually adjusted assuming spatial isotropy in the search of preferential directions.

### 2.2. Supervised machine learning techniques for target characterization

Recent advances in computer science offer the possibility to couple machine learning with traditional statistical methods such as Bayesian networks (Benavoli et al., 2017). Bayesian networks have shown their potential in problem domains with manifold variables of different typologies, where the medical and occupational health domain is a showcase of their performance. Concretely, information theory in combination with Bayesian networks is used to respond to the different stages of this study, quantifying the reduction of uncertainty brought by each medical variable to the knowledge of the health state.

On this basis, once the Bayesian model is built as a result of the machine learning process aimed to discover significant relationships in the problem space search, the Kullback-Leibler (KL) divergence is used as a measure of strength in the relationship between two nodes that are directly connected by an arc (Conrady et al., 2015). This parameter allows measuring how the probability distribution in each variable drifts away from

the state of health (target node). From a mathematical viewpoint, let P and Q represent the distribution of two joint probabilities defined for the same set of variables or X nodes.

$$D_{KL}\big(P(X)||Q(X)\big) = \sum_{X} P(X) log_2 \frac{P(X)}{Q(X)} \quad (1)$$

For target node characterization, the relative weight value is shown as a fraction of the maximum KL Divergence value. Likewise, these weights can be depicted as the global contribution percentage of each arc to the target node quantifying the value between two directly connected nodes $D_{KL}$(Parent|Child) and the sum of all KL Divergence values across the network. In addition, the independence test G is computed from the KL divergence of the relationship, thus its value is reckoned from the network.

## 3. Results

Given the need to clarify the understanding of occupational health risks triggering workers' sick leave, this section presents the preliminary results obtained from the application of Bayesian machine learning and geostatistics to the occupational health data for the four service activities under analysis. The results outline the findings obtained based on the four methodological levels summarised in Figure 1.

In the first place, a general Bayesian network was built delving into the statistical association stemming from the state of health and each variable in the model, considering all service activities (administrative and auxiliary services, financial and insurance services, education, and hostelry). From the resulting Bayesian network, a relationship analysis was carried out. The more representative parent-child connections were identified. These relationships are shown in Table 1, where *age* excels by its high impact, followed by the *location* and the *total cholesterol*.

Table 1 – Characterization of the target node (health state). Relationship analysis for the most representative medical, socioeconomic and bioclimatic variables.

| Parent | Child | KL(Parent\|Child) | Relative weight | Contribution | G Test |
|---|---|---|---|---|---|
| Health state | Age | 0.0521 | 1.0000 | 16.6440% | 5,374.1635 |
| Health state | Location | 0.0341 | 0.6552 | 10.9056% | 3,521.2995 |
| Health state | Total Cholesterol | 0.0287 | 0.5504 | 9.1604% | 2,957.8012 |
| Health state | Drug Consumption | 0.0213 | 0.4084 | 6.7969% | 2,194.6465 |
| Health state | Hearing test | 0.0169 | 0.3238 | 5.3896% | 1,740.2371 |
| Health state | Spine Observation | 0.0149 | 0.2852 | 4.7465% | 1,532.5901 |
| Health state | Limbs Observation | 0.0145 | 0.2789 | 4.6413% | 1,498.6285 |
| Health state | Physical Limitations | 0.0126 | 0.2416 | 4.0214% | 1,298.4567 |
| Health state | Minimum Rainfall | 0.0100 | 0.1922 | 3.1997% | 1,033.1395 |
| Health state | Population | 0.0074 | 0.1417 | 2.3587% | 761.5890 |
| Health state | Annual Rainfall | 0.0073 | 0.1410 | 2.3476% | 758.0027 |
| Health state | Sleep Quality | 0.0068 | 0.1312 | 2.1836% | 705.0538 |
| Health state | Maximum Temperature | 0.0067 | 0.1280 | 2.1306% | 687.9423 |

In the second place, four supervised Bayesian networks were built, corresponding to each of the four defined service activities and whose common target node was the health status of the worker. The application of a Naïve Bayes algorithm allowed to generate a pragmatic network structure for the analysis of the influence of

each variable on the health status of the workers (Figure 2). The characterization of the target node revealed that age, location, and total cholesterol, previously identified as the most significant factors in the general network of the service sector, also present a high impact on all the concrete service activities under study. In that context, the authors have considered the need to deepen the understanding of those variables that are a priori not that significant, but which may hold key differentiating aspects within each population group.

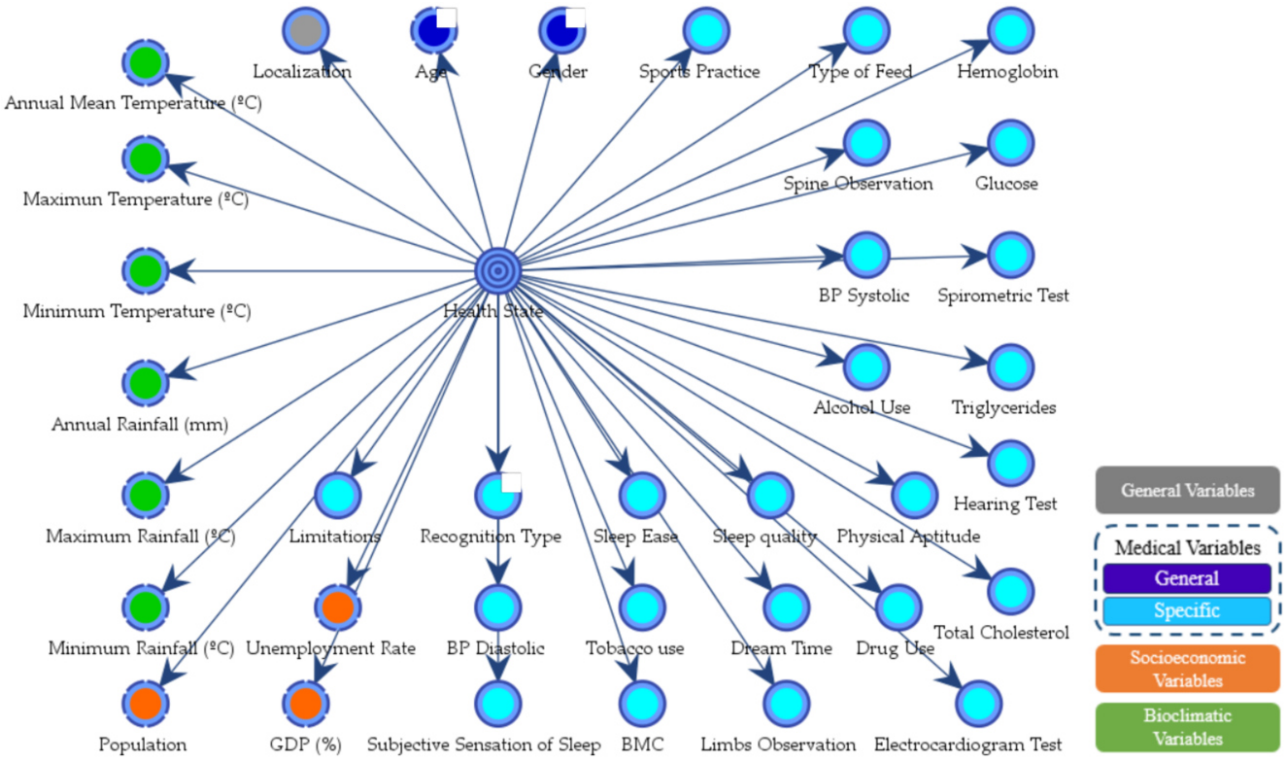


Figure 2 – Supervised Bayesian network built with Naïve Bayes algorithm. The graph presents the administrative and auxiliary services network with the target node (health state) in the center.

When looking at the distribution of contributions of each variable to the characterization of the state of health, it is found that the nervous system (15%-19%) matches to a high extent the characterization of the medical examinations of healthy workers (64%-70%). The most significant medical conditions conditioning these two states are age, total cholesterol, and location, whereas hearing problems and drug use are always reflected as differential variables. As an example, after an inference analysis on patients with high levels of total cholesterol belonging to hostelry services, a greater impact could be seen on elderly workers (>50) belonging to the autonomous community of the Basque Country (38.26 % of registered cases) located in the North of Spain. In contrast, it can be concluded the strong need to provide a higher level of granularity on the musculoskeletal (8%-11%) and cardiovascular (6%-9%) pathologies, as here the differences among possible additional differential variables, even if relevant from a mathematical point of view, they are not meaningful from a policy perspective (Table 2).

The great horizontality of variables such as *age, location,* and *total cholesterol* directed this study towards the need to add value to those differentiating variables of the musculoskeletal and cardiovascular systems. This situation leads to the spatial representation of the variable's *spine observation, annual precipitation (BIO 12), limbs observation, and annual mean temperature (BIO 1)* under an ordinary kriging approach (Figure 3). This approach allows identifying both a spatial distribution of spinal problems and potentially related extremities, as well as two clearly differentiated regions where these problems have a higher impact. Particularly, in the Northeast

of Spain, except Catalonia, and the South, with vascular problems such as the presence of varicose veins. As for the Western part of Spain, a higher rate of spinal problems, derived from muscle contractures or other minor discomforts, is identified. Based on Bayesian results, it can be demonstrated that this type of injury is related to a great extent to pathologies of the musculoskeletal system which is potentially present in service activities such as hospitality (31.23%) and administration (32.05%). In addition, it is also seen how these pathologies are also an underlying cause for the appearance of problems in the extremities.

Table 2 – Local impact analysis of data over the target node for the states representing musculoskeletal and cardiovascular systems by service activity.

| | **Musculoskeletal System** | | **Cardiovascular System** | |
|---|---|---|---|---|
| | Relative Binary Mutual Information | | Relative Binary Mutual Information | |
| Group of Work | Spine Observation | Annual Rainfall | Limbs Observation | Annual Temperature |
| Administrative and auxiliary services | 3.4737% | 0.64% | 1.7552% | 1.3914% |
| Financial and insurance activities | 2.2066% | 0.5638% | 1.5169% | 0.9984% |
| Education | 3.6072% | 0.7496% | 0.9042% | 0.4487% |
| Hostelry | 3.2213% | 1.3584% | 2.1152% | 0.0360% |

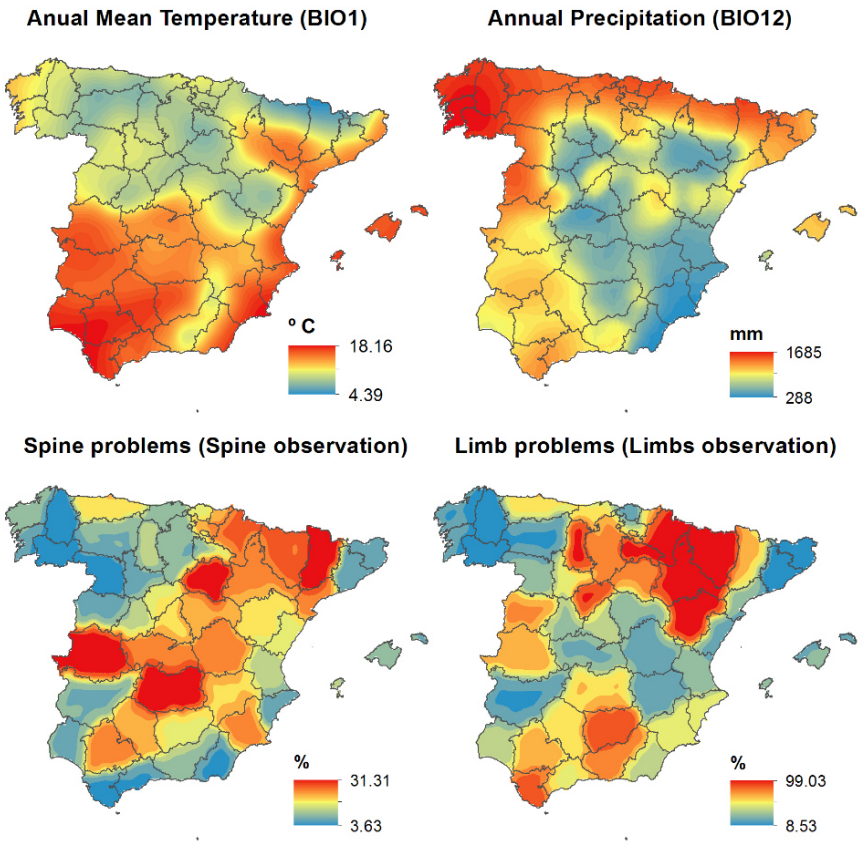


Figure 3 – Distribution maps for annual mean temperature (ºC), annual rainfall (mm), and spine and limb observation variables using rate data between 2012 and 2016 interpolated by Ordinary Kriging.

## 4. Discussion and Conclusions

Given the greater complexity of characterizing the musculoskeletal and cardiovascular systems, and the impossibility of achieving greater conceptual discrimination of the detailed variables with Bayesian networks, a new level of granularity is needed. Here, an ordinary kriging approach enters into play, offering the possibility to differentiate and obtain meaningful policy findings at a regional scale, revealing what are the exact implications, above all, of the bioclimatic variables and how they affect unhealthy workers within the service sector.

As a showcase of the potential of this combined approach, Figure 4 shows the Bayesian results of the inference analysis on the medical variables *spine* and *limbs observation*, and the bioclimatic variables *annual mean temperature (ºC) (BIO1)* and *annual rainfall (mm) (BIO12).* This inference is carried out with two variables of reference that are the service sector activities (group of work) and the state of health (pathology). In general, the results allowed to conclude a greater impact of spinal problems for hostelry activities, as well as a direct relationship of this pathology with limb problems, increasing the cases of workers with adverse vascular conditions by 19.64%. This is an example, which shall be complemented always with a more granular spatial mapping to deepen on the regional variations.

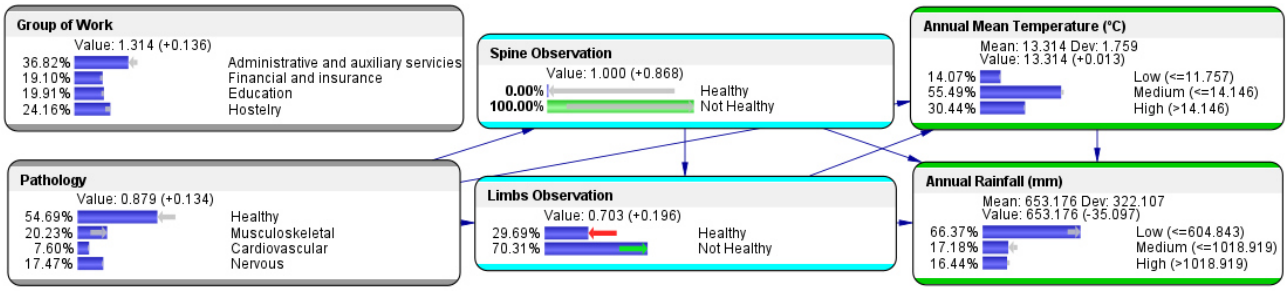


Figure 4 – Inference results for the work groups when the evidence reflects spinal problems.

In conclusion, the results of this study revealed that variables such as age, location, and cholesterol, with contributions to the general network between 9-17%, are generally critical for the characterization of the health status of workers in the service sector. To a second extent, it was possible to identify a series of differentiating variables such as pine observation, annual precipitation (BIO 12), limbs observation, and annual mean temperature (BIO 1) that despite not being extremely significant from a mathematical point of view, they play a key role and show a great impact at regional level.

Likewise, this article exposes the potentialities of the combination of Bayesian machine learning complemented by geostatistics to translate the complex occupational health problem of workers' health status into evident visual findings that can feed medical policy developments across different service activities. At this stage, it is already possible to demonstrate the high influence of bioclimatic and socioeconomic variables within the medical decision making of a worker health state. Looking forward, further analysis is needed to identify more health risk factors that can be derived, for example, from the impact of high temperatures or income level. In this respect, depending on the data available and the scope of the analysis, more sophisticated geostatistical approaches would have also to be explored.

## References

A. Benavoli, A., G. Corani, J. Demsar, M. Zaffalom. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. Journal of Machine Learning Research. 2017.

A. Orlov, J. Sillmann, K. Aunan, T. Kjellstrom, and A. Aaheim, "Economic costs of heat-induced reductions in worker productivity due to global warming," Glob. Environ. Chang., vol. 63, no. September 2019, p. 102087, 2020, doi: 10.1016/j.gloenvcha.2020.102087.

B. Y. Wondmagegn et al., "Increasing impacts of temperature on hospital admissions, length of stay, and related healthcare costs in the context of climate change in Adelaide, South Australia," Sci. Total Environ., vol. 773, p. 145656, Jun. 2021, doi: 10.1016/J.SCITOTENV.2021.145656.

C.-H. Chang, R. Shao, M. Wang, and N. M. Baker, "Workplace Interventions in Response to COVID-19: an Occupational Health Psychology Perspective," *Occup. Heal. Sci.*, vol. 5, no. 1–2, pp. 1–23, Mar. 2021, doi: 10.1007/S41542-021-00080-X/TABLES/1.

Eurostat, "Contributions of each sector - Institutional sector accounts - Eurostat". European Commission, 2022. https://ec.europa.eu/eurostat/web/sector-accounts/detailed-charts/contributions-sectors

J. B. Awotunde, A. E. Adeniyi, R. O. Ogundokun, G. J. Ajamu, and P. O. Adebayo, "MIoT-Based Big Data Analytics Architecture, Opportunities and Challenges for Enhanced Telemedicine Systems," *Stud. Fuzziness Soft Comput.*, vol. 410, pp. 199–220, 2021, doi: 10.1007/978-3-030-70111-6_10.

K. L. Ebi et al., "Extreme Weather and Climate Change: Population Health and Health System Implications," https://doi.org/10.1146/annurev-publhealth-012420-105026, vol. 42, pp. 293–315, Apr. 2021, doi: 10.1146/ANNUREV-PUBLHEALTH-012420-105026.

P. Goovaerts. Geostatistics for Natural Resources Evaluation. Applied Geostatistics Series. Oxford University Press, New York, NY (USA), 837 483 p., 1997.

S. Conrady, L. Jouffe. Bayesian Networks & BayesiaLab - A Practical Introduction for Researchers. Bayesia USA. 2015. ISBN-10: 0996533303.

S. Gerassis, C. Boente, M.T.D. Albuquerque, M.M. Ribeiro, A. Abad, J. Taboada, "Mapping occupational health risk factors in the primary sector—A novel supervised machine learning and Area-to-Point Poisson kriging approach" *Spatial Statistics*, vol. 42, 100434, 2021.