Mathematical Geosciences
https://doi.org/10.1007/s11004-023-10087-5

SPECIAL ISSUE

Unpacking Occupational Health Data in the Service Sector: From Bayesian Networking and Spatial Clustering to Policy-Making

María Pazo[1] · Carlos Boente[2] · Teresa Albuquerque[3,4,5] · Saki Gerassis[1] · Natália Roque[3,4] · Javier Taboada[1]

Received: 1 February 2023 / Accepted: 8 July 2023

Abstract

The health status of the service sector workforce is a significant unknown in the field of medical geography. While spatial epidemiology has made progress in predicting the relationship between human health and the environment, there are still important challenges that remain unsolved. The main issue lies in the inability to statistically determine and visually represent all spatial concepts, as there is a need to cover a wide range of service activities while also considering the impact of numerous traditional medical variables and emerging risk factors, such as those related to socioeconomic and bioclimatic factors. This study aims to address the needs of health professionals by defining, prioritizing, and visualizing multiple occupational health risk factors that contribute to the well-being of workers. To achieve this, a methodological approach based on the synergy of Bayesian machine learning and geostatistics is proposed. Extensive data from occupational health surveillance tests were collected in Spain, along with socioeconomic and bioclimatic covariates, to assess potential social and climate impacts on health. This integrated approach enabled the identification of relevant patterns related to risk factors. A three-step geostatistical modeling process, including variography, ordinary kriging, and G clustering, was used to generate national distribution maps for various factors such as annual mean temperature, annual rainfall, spine health, limb health, cholesterol, age, and sleep quality. These maps considered

✉ María Pazo
maria.pazo@uvigo.gal

1 GESSMin Research Group, Department of Natural Resources and Environmental Engineering, University of Vigo, Lagoas Marcosende, 36310 Vigo, Spain

2 CIQSO-Center for Research in Sustainable Chemistry, Associate Unit CSIC-University of Huelva "Atmospheric Pollution", Campus El Carmen s/n, 21071 Huelva, Spain

3 Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal

4 Centro de Estudos de Recursos Naturais, Ambiente e Sociedade (CERNAS) - Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal

5 ICT| Universidade de Évora, Largo dos Colegiais 2, 7000 Évora, Portugal

Published online: 07 August 2023

four target activities—administration, finances, education, and hospitality. Remarkably, bioclimatic variables were found to contribute approximately 9% to the overall health status of workers.

Keywords Health data · Information theory · Bayesian learning · Ordinary kriging · *G* clusters

1 Introduction

The service sector, generally referred to as the tertiary sector of the economy, includes the provision of services to other businesses, including final consumers. In the European Union (EU), services account for approximately 70% of the Union's gross domestic product (GDP) and employment (Eurostat 2022). In some countries, such as the United States, it could be as high as 80% (World Bank Group 2022). Some of the most common areas of the service sector are tourism (e.g., accommodation, and travel agents), hospitality (e.g., food services), education, real estate, transport, and banking. The wide range of activities available in this sector makes it extremely difficult to assess the health status of their workforce. For example, the hospitality industry offers employment opportunities to minority groups, such as immigrants, women, or youth with low educational attainment. In many cases, these activities are characterized as labor-intensive and are related to long working hours and a high workload (Rydzik and Anitha 2020; Xu et al. 2020).

The impact of COVID-19 has exacerbated this situation (Chang et al. 2021). Globally, projections of investment in the health of the workforce appear inadequate, which undermines the future sustainability of the workforce and health systems (World Health Organization 2016). In practice, the tasks of men and women are often different, which creates health risks at work for each gender. Despite the advances carried out by medical geography and spatial epidemiology to predict spatial patterns of disease incidence, the abundance and accuracy of occupational health risk maps are still very limited (Gerassis et al. 2021). The introduction of geostatistical modeling to support occupational health decision-making is relatively new. Notably, Dos Santos et al. (2020) used a geostatistical wave model to determine healthy workspaces for rural workers exposed to tractor noise. Other application examples include the use of Bayesian geostatistical binary regression to model the 2-week disease prevalence rate among workers (Wen et al. 2021) or the geostatistical analysis of mental health in construction workers (Yuvaraj and Thulasimala 2022). While significant progress is being achieved, it remains a challenge how best to combine big data from occupational health with data from other domains to examine the relationships between workers and their environments. This is partly due to the multiple variables to be represented without a holistic approach to unify the field of the problem, and even more, to discover these differentiating variables.

Medical geography is cross-disciplinary (Jerrett et al. 2010). In practice, medicine has been integrated into a range of disciplines, including sociology, economy, history, ecology, biology, anthropology, and political science. The geography of health illustrates the importance of medical geography by identifying geographic variations

within health and healthcare systems. Health geography has evolved from medical geography in recent decades, and the construction process continues (Moon 2020). In this study, medical and health geography is developed in occupational health to support public health policy and planning in the service sector. Here, the impact on the health of climate change (Orlov et al. 2020) is understood to be a key consideration.

Climate change is already influencing the intensity, severity, and frequency of heat waves (Perkins-Kirkpatrick and Lewis 2020), which points to increased cardiovascular and respiratory diseases and, subsequently, mortality, particularly during extreme heat events (Ho et al. 2015). Rising temperatures are expected to open the door to a growing number of diseases in the coming years whose effects could be worse at work. This could have a major impact on service sector performance, as related surveys reveal effects on morbidity, reduced productivity of individuals, and increased sick leave (Ebi et al. 2021; Wondmagegn et al. 2021). At present, an important question remains unanswered: what is the real impact of climate variables on the health of the service sector? To clarify, annual mean temperature and annual rainfall were introduced as covariates, enabling insight into the effect of location and environmental exposure on the workers' health.

The concrete goal of this study is to introduce a methodological decision-to-visualization process to understand and measure the occupational health risk factors, together with local climatic conditions, leading to unhealthy workers that may be unfit to perform their duties. This is achieved by taking advantage of the latest advances in probabilistic models of structural equations (PSEMs) using Bayesian modeling and geostatistics. The occupational health data were obtained from the annual workers' medical checks, socioeconomic covariates were obtained from the Spanish national agencies, and bioclimatic variables from the WorldClim database (Fick et al. 2017; Panagos et al. 2017). This survey involved research of big data using machine learning techniques, interpreted within a framework of spatial organization, aimed at supporting health policy development and targeting occupational strategies of disease monitoring at work. By putting theory into practice, a progressive learning approach is proposed to build a methodological process that leads to informed decision-making. Bayesian networks were selected as a proxy to manage the large number of variables associated with this type of complex problem. Specifically, Bayesian machine learning was used to develop a PSEM where the health status is the target node of the model. The use of a hierarchical Bayesian structure allows one to obtain a compact representation of the probabilistic dependencies between the multiple variables used to characterize the health status (Gao et al. 2022; Njah et al. 2021; Letta et al. 2022). Importantly, the Bayesian equation model offers the possibility of inserting latent variables into the structure, facilitating the identification of new relationships between variables and, consequently, reducing the complexity of the problem (Peterson et al. 2020; Keter et al. 2022).

The introduced methodological approach is understood as a unified and renewed conceptualization of a series of prior works. Modeling geospatial uncertainty with Bayesian models, including advanced deep learning techniques, is something already extended in the literature (e.g., Hoffimann et al. 2022; Kirkwood et al. 2022). Structural equation models (SEMs) have been an essential tool for causal analysis in social and behavioral sciences for more than 50 years (Pearl 1998).

Combining both domains leads to the PSEMs recently popularized by Conradi and Jouffe (2015) with the recent scale-up of dedicated software for automated machine learning (AutoML) both in research and industrial applications. Furthermore, mapping is a precise way of simplifying reality (Lahr and Kooistra 2010; Albuquerque et al. 2017); however, a two-dimensional representation may only aggregate and display a limited number of attributes. Consequently, when examining complex scenarios, such as environmental or epidemiological characterization, there is a need to reduce the dimensionality of the problem. Risk maps, which are widely cited in the literature, are highly relevant to the visualization of spatial models, and powerful tools to support policy development within a complex risk assessment framework. Examples of these practices include the distribution of pollutant levels or vulnerability assessment. Geostatistical techniques are based on the theory of regionalized variables (Matheron 1971) whereby variables within a region have spatially structured and random properties (Journel and Huijbregts 1978). Geostatistics is based on an extensive methodological approach and goes beyond the simple development and application of mathematical (probabilistic) models and methods. A key challenge is to analyze the practical problems to be solved and to formalize them in terms of concepts. In anticipating risk, it is imperative to emphasize the appropriateness of the likelihood that future estimated values will exceed the maximum permissible values. The delineation of zones of high and low impact requires the interpolation of the selected covariates to the nodes of a regular grid, making possible the assessment and prediction of prevalent spatial patterns, as guidance to a more sustainable management (Goovaerts 1997).

In that manner, the added value is the possibility to identify and characterize those variables that may have a differentiating impact that is not meaningful from a mathematical point of view (Kiebish et al. 2020; Mohamed et al. 2021). 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's health status will no longer be subject to only a series of traditional medical tests and underlying medical conditions (Awotunde et al. 2021).

The remainder of the manuscript is organized as follows. Section 2 explains the methodology employed to develop the PSEM based on a Bayesian hierarchical structure, as well as the probabilistic induction of latent factors. In addition, a preliminary discussion on data description is currently under consideration. The methodological approach to geostatistical modeling is also discussed in this section. Section 3 shows the results of the PSEM and the spatial representation of the most significant variables identified to characterize the health status of service sector workers, namely, annual mean temperature, annual rainfall, spine health, limb health, cholesterol, age, and sleep quality. Finally, Sect. 4 argues how the use of the PSEM improves the interpretability of complex problems, its combination with geospatial modeling being a key tool for health decision-making.

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 and 2016 throughout the Spanish 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 hospitality (15,611). Each clinical examination was conducted in compliance with Spanish occupational health legislation (Ley 31/1995). Data were anonymized and released only after a period which does not interfere with processes performed by the relevant occupational health organizations and hospital services conducting the medical tests and gathering major information about the state of workers' health. Health status is defined by the major health risk factors underlying the disease, including key physical conditions and health patterns. Importantly, this study goes 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 in the future. The WorldClim database was used for bioclimate data extraction (Fick et al. 2017), and socioeconomic data were obtained from the National Statistical Institute (INEbase). Figure 1 in Sect. 2.2 (Analytical Steps) and Table 1 in Sect. 3 (Results and Discussion) give a comprehensive overview of the medical, bioclimatic, and socioeconomic variables used.

2.2 Analytical Steps

Procedurally, this research is conducted through a five-level approach, as illustrated in Fig. 1. First, an unsupervised Bayesian network is constructed to uncover direct probabilistic relationships between the initial 48 manifest variables (level 1). Secondly, latent class modeling is introduced to define representative subgroups for analysis (level 2). Thirdly, the probabilistic structural equation model (PSEM) is developed, in which latent factors and target variable health status provide an overall representation of the field of study (level 3). Later, for each activity group, the health status also acts as a target node for which the relevant patterns, associated with the type of work performed, are ascertained (level 4). These four levels are associated with the development of a Bayesian methodology for which BayesiaLab v.10.2 (www.bayesialab.com) was used. Finally, at level 5, a three-stage geostatistical approach to the computation of distribution maps was conducted, to deepen the network's knowledge from a spatial point of view. Ordinary kriging, followed by a *G*-group analysis, was used in the four service activity groups under analysis. ArcGIS v-10.2.2 and SpaceStat v-4.1.26 were used for computation.

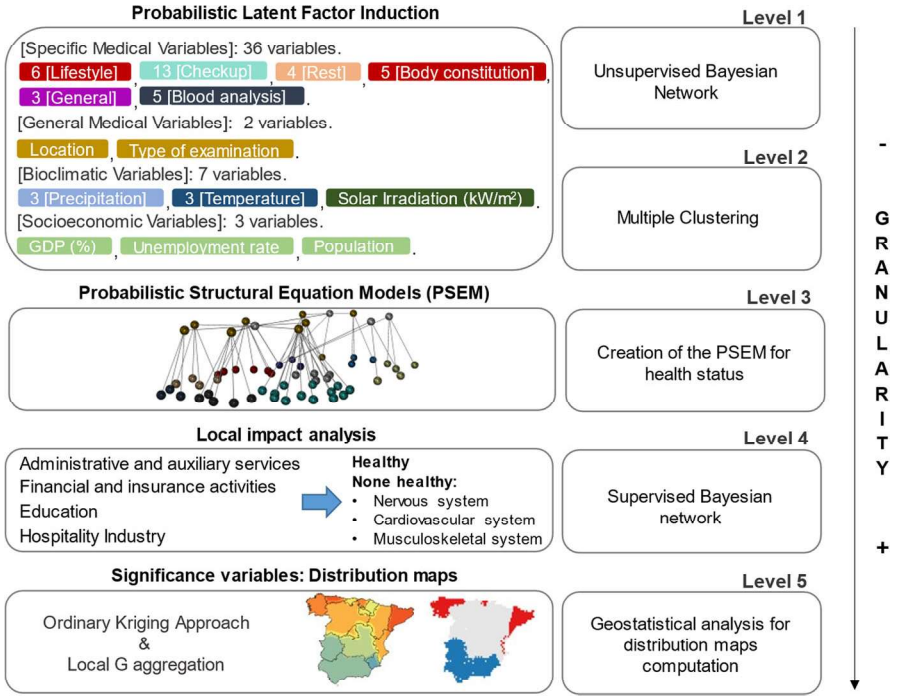

Fig. 1 The methodological process was implemented with five levels of analysis. The colors shown in the description of the probabilistic latent factor induction represent each cluster of variables identified and the corresponding number of variables analyzed

2.3 Probabilistic Latent Factor Induction

This section introduces levels 1 and 2 of the analysis, which aim to exploit and interpret complex data by capturing direct probabilistic relationships and latent subgroups within the dataset:

1. *Level 1*: Induction of latent factors (unobserved variables) begins with the development of an exploratory Bayesian network (Pearl 1988). For this purpose, an unsupervised network is built to represent the strongest probabilistic relationships that exist between the manifest or observed variables under analysis. This approach has great potential to create a global framework that reveals hidden trends to healthcare providers. In a formal sense, the joint distribution $p(x)$ of a random set of variables $X = (X_1, \ldots, X_m)^T$ may be described as follows (Murphy 2012)

$$p(x) = p(x_1, \ldots, x_m) = \prod_{i=1}^{m} p(x_i | x_{\pi i}), \tag{1}$$

which denotes that $x = (x_1, \ldots, x_m)^T$ is a realization of X, and $x_{\pi i}$ represents a realization of parent variables $X_{\pi i}$ of each X_i.

Table 1 Results of cluster analysis

Cluster	Variables	Node force
Precipitations	Minimum precipitation, maximum precipitation, and annual precipitation	4.7659
Temperatures	Minimum temperature, maximum temperature, annual temperature	4.0492
Body constitution	Age, gender, height, weight, body mass index (BMI)	3.9049
Checkup	Blood pressure (BP) diastolic, BP systolic, electrocardiogram test, hearing test, limb test, lung auscultation, neurological condition, spine test, spirometric pattern, vision test, cardiovascular rate, lung rate, skin, and mucosal test	3.2048
Blood analysis	Glucose, hematocrit, hemoglobin, total cholesterol, and triglycerides	2.7278
Rest	Hours of sleep, sleep quality, start of sleep, subjective feeling sleep	2.4094
General	Aptitude, physical limitation, patient segmentation	1.4003
Socioeconomic	GDP (%), population, unemployment rate	0.9238
General medical	Localization, type of recognition	0.5945
Lifestyle	Alcohol use, drug use, sports practice, tobacco use, type of food, and other uses	0.3807

At this stage, workers' health status is excluded from the learning process for variable clustering. The reason is that the worker's health status will be used as a target variable in the conclusion of the PSEM; therefore, it is not relevant to add it to the local learning rather than to analyze its liaison with those relevant clusters of variables.

2. *Level 2*: The goal is to define the best-fit clusters of variables, for which an agglomerative hierarchical clustering algorithm supported in BayesiaLab v.10.2 is used. Specifically, arc force between manifest variables is used as a probabilistic measure to gradually group highly correlated variables (Conrady and Jouffe 2015). On this basis, once the global Bayesian model is built, because 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. Under a formal approach, allow P and Q to represent the distribution of two common probabilities defined for the same set of variables or X nodes (van Erven and Harremos 2014).

$$D_{\mathrm{KL}}(P||Q) = \sum_{x \in X} P(x) \log_2 \frac{P(x)}{Q(x)}. \tag{2}$$

Once the variables were grouped, a latent class model was introduced to define representative subgroups for the analysis. For this purpose, a naive architecture

was created, in which the factor variable is the parent of the manifest variables. The established latent factors provide a homogeneous, compact, and stable representation of the local joint probability distributions (JPDs) defined by the associated manifest variables. Mathematically, an expression of a latent cluster model is provided by

$$P_{Yi} = \sum_{n=1}^{N} P_{Xn} P_{(Yi|Xn)}, \tag{3}$$

where P_{Xn} describes the probability that an observation from the set of observed variables $(Y_1, \ldots, Y_n)$ describes the latent variable X, which belongs to a latent class $(n = 1, 2, ..., N)$. On the other hand, $P_{(Y_i|X_n)}$ is the conditional probability of getting an observation with a response pattern $Yi = (y_1, ..., y_n)$, while belonging to a class n of a latent variable X (Yousefi and Tucker 2022). Additionally, to characterize the probabilistic relationships between the latent factor and its manifest variables an expectation–maximization (EM) algorithm is used. This criterion for estimating the maximum likelihood permits us, through the Bayes theorem, to obtain the subsequent probability of an observation belonging to a given class n.

$$P_{(Xt|Yi)} = P_{(Yi|Xn)} \frac{P_{(Xn)}}{P_{(Yi)}}. \tag{4}$$

2.4 Probabilistic Structural Equation Modeling (PSEM)

Subsequently, levels 3 and 4 involve the final construction of the Bayesian network using supervised health condition analysis. Once the observed variables have been linked to the underlying factors, the excluded target node (health status) mentioned in Sect. 2.3 is incorporated into the final network structure to finalize the probabilistic structural equation model (PSEM).

Specifically for level 3, the relationships between factors, marginal variables, and the target node were established based on the combination of the presented probabilistic theory and expert criteria. To characterize the target node, the relative weight value is displayed as a fraction of the maximum KL divergence value. Similarly, these weights can be represented as the global contribution percentage of each arc to the target node, quantifying the value between two directly connected nodes DKL (parentlchild), and the sum of all KL divergence values across the network. This analysis enables the identification of clusters that have the greatest impact on the health of workers and enables the implementation of targeted interventions and strategies to improve their well-being. The opportunity to introduce expert criteria opens the door to a more flexible approach where different probabilistic configurations can be studied.

To validate the PSEM, the contingency table fit (CTF) metric was used to measure the quality of the representation of the joint probability distribution via the latent variable. BayesiaLab's CTF is defined as the entropy value (H_B) of the created network B compared to the value H_C of the fully connected network C. In addition, H_U represents

the entropy value of the data considering the disconnected network U (BayesiaLab, n.d.)

$$C_B = 100\frac{H_U(D) - H_B(D)}{H_U(D) - H_C(D)}. \tag{5}$$

Therefore, if the current system can generate a precise representation of the fully interconnected joint probability distribution, the CTF will assume a value of 100. Conversely, if the network represents a completely disconnected structure where all variables are marginally independent, the CTF will be equal to 0.

2.5 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 (Abad et al. 2019; Gerassis et al. 2019). Concretely, information theory in combination with Bayesian networks can be used to respond to the different stages of this study, allowing one to quantify the reduction of uncertainty brought by each medical variable to the knowledge of the health state.

Accordingly, a set of supervised networks corresponding for each working group is established. This higher degree of granularity, in which health has acted as a target node, reveals trends associated with specific activities performed by workers. The relative mutual information value is employed at this level to quantify how much information a variable provides (X_n) knowledge about the characterization of the patient's health status (X_T). Mathematically, the formula used in the calculation was as follows

$$I_R(X_T, X_n) = \frac{I(X_T, X_n)}{H(X_T)} = \frac{H(X_T) - H(X_T|X_n)}{H(X_T)}. \tag{6}$$

2.6 Spatial Representation

Spatial models of selected attributes, annual mean temperature, annual rainfall, spinal health, limb health, cholesterol, age, and quality of sleep have been constructed using three-step geostatistical modeling:

1. Experimental isotropic variograms were computed, and theoretical models were fitted.
2. Ordinary kriging (OK) was used as an interpolating algorithm for the original rank values.
3. Finally, local G clustering (Getis and Ord 1992) allowed us to measure the degree of association that results from the concentration of weighted points (or region represented by a weighted point), and all other weighted points included within

a radius of distance from the original weighted point. Considering a given zone divided into n regions, $I = 1, 2, \ldots, n$, where each neighbor is distinguished by a point for which the Cartesian coordinates are known. Each I has associated with it a value x (a weight) taken from a variable X. The variable has a natural origin and is positive. The statistic of $G(i)$ developed below allows us to test the assumptions on the spatial concentration of the sum of the values x associated with the points j in d of the ith point. The following statistical information is obtained

$$G_i(d) = \frac{\sum_{j=1}^{n} W_{ij}(d)x_j}{\sum_{j}^{n} x_j}, \quad j \text{ not equal to } i,$$

where W_{ij} is a symmetric one/zero spatial weight matrix with ones for all links defined as being within distance d of a given i; all other links are zero, including the link of point i to itself. The numerator is the sum of all xj inside of i but without including xi. The divisor is the sum of all xj, except xi.

3 Results and Discussion

This section presents the results obtained from a PSEM based on a Bayesian machine learning construct to the subsequent spatial assessment using a geostatistical methodology. The results describe the findings in identifying the most significant occupational health risks.

3.1 Probabilistic Latent Factor Induction

At the first level, an unsupervised overall model was created from the initial 48 marginal variables, excluding the worker's health status. The main purpose of the established network was to find clusters in terms of probabilistic relations between nodes. Furthermore, as the construction of an unsupervised network is the first step of a PSEM, it was necessary to set a maximum number of variables per cluster. Specifically, to get an understandable representation of the domain, 10 initial clusters were selected with groups between 2 and 13 variables. This process was carried out with a hybrid approach; that is, based on the algorithmic clustering proposal, the expert criterion was introduced for an adjusted model of the analysis scenario. Finally, Table 1 provides the results of the cluster analysis, including the computed node strength value for each cluster. The node force value represents the sum of the arc forces of all interconnected arcs to the cluster, providing a quantitative measure of the individual influence of each cluster for the domain under analysis.

The clustering variables of the initial Bayesian model are aggregated into a higher-level subnetwork corresponding to the four main clusters under study: specific medical variables, socioeconomic variables, bioclimatic variables, and general medical variables. Multiple clustering algorithms within BayesiaLab create a discrete factor for every subset of grouped marginal variables. In this case, the optimum number of states to represent the JPD of the marginal variables is automatic. The only limitation, as

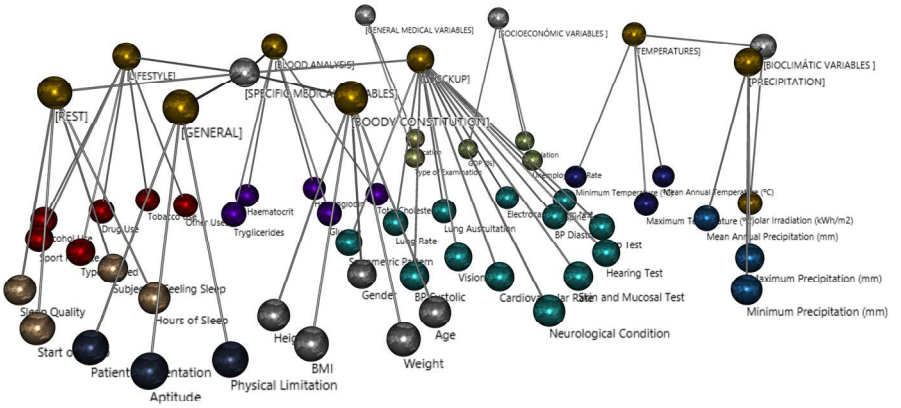

Fig. 2 Multiple Bayesian models were generated from a data clustering analysis. The colors of the marginal nodes are established following the groups obtained after the variable group analysis. The induced latent factors for each group of marginal variables are shown in the top layer in yellow, while the main latent factors are shown in gray

recommended by Conrady and Jouffe (2015), is to set the number of classes from 2 to 5. The spatial representation of the domain is depicted in Fig. 2.

3.2 Probabilistic Structural Equation Model (PSEM)

Once the multiple clustering is completed, the PSEM design shows four Bayesian subnetworks surrounded by the induced latent factors and their marginal variables (Fig. 3). To analyze the influence of all latent factors on workers' health conditions, the connection between health status (target node) and the four main clusters under study (specific medical variables, socioeconomic variables, bioclimatic variables, and general medical variables) was manually assigned (expert criteria).

Based on the PSEM, the statistical association between health status and each cluster of the model was further investigated. The most representative parent–child relationships are shown in Table 2, which accounts for all service activities (administrative and ancillary services, financial and insurance services, education, and hospitality) in the analysis.

The cluster of specific medical variables provides more insight into the target node than the rest of the latent cluster (80%). Likewise, general medical variables provide 10% of the knowledge required to characterize the health status of workers. From the point of view of predictive significance, the marginal variable workplace (location) contributes 96.6820% to the reduction of the uncertainty of the latent cluster in which it is located, compared to 1.0118% of the recognition type node. The high significance of the location variable within the latent cluster defined as general medical variables highlights the importance of providing spatial patterns that represent how the variables with the greatest impact on the health status of workers in the tertiary sector are distributed. Lastly, the authors consider that it is very important to emphasize the importance of bioclimatic variables on workers' health indicators. Obtaining a contribution of nearly 9% highlights the real influence of climate.

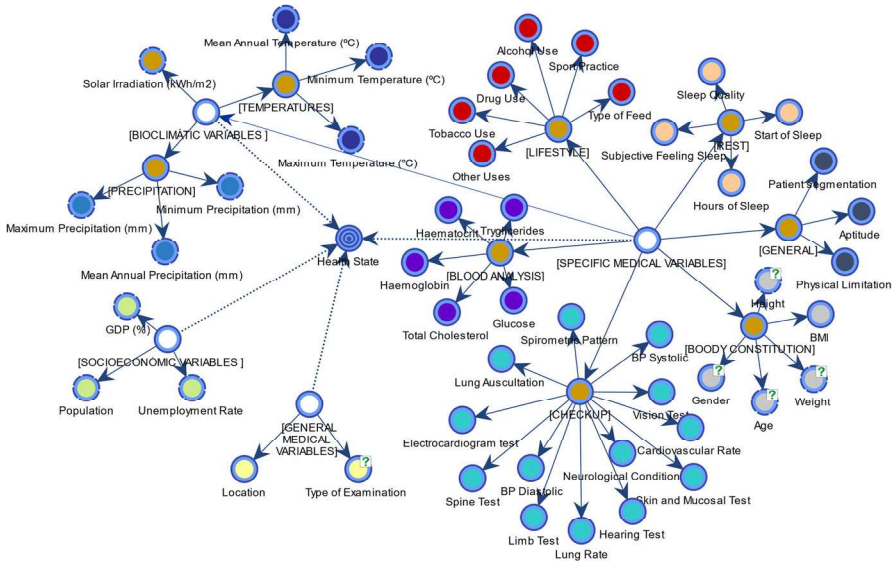

Fig. 3 Final PSEM structure with two distinct levels of complexity built using a semi-supervised Taboo algorithm

Table 2 The parent–child relationship and the analysis of the contribution between the clusters of the model on the characterization of the health status of workers in the service sector

Parent	Child	KL (parent|child)	Relative weight	Contribution (%)
Health state	Specific medical variables	0.0531	1	79.499
Health state	General medical variables	0.0067	0.1261	10.026
Health state	Bioclimatic variables	0.0011	0.1108	8.807
Health state	Socioeconomic variables	0.0210	0.0210	1.668

To identify the most dominant variables in terms of the latent factor uncertainty, the relative mutual information (RMI) of the arcs was computed exclusively between the manifest latent factors and the target node health state (Fig. 4). The results show that the cluster of specific medical variables is the factor that most reduces uncertainty about having a healthy or unhealthy patient by an average of 5.3556%. For clusters of bioclimatic and general health variables, the results are similar with percentages of 0.8840% and 0.7298%, respectively. Social and economic variables have the lowest MI values.

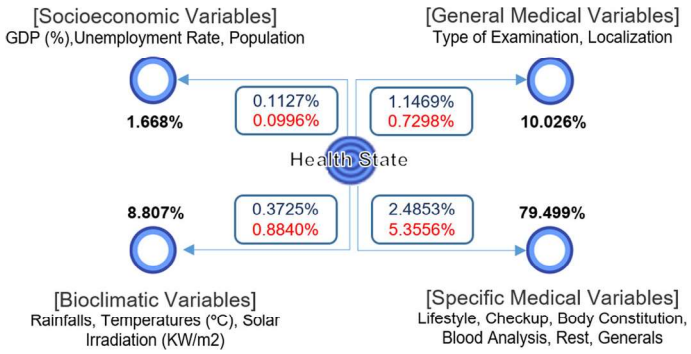

Fig. 4 PSEM arc's mutual information (RMI child color blue and RMI parent color red) and contribution visualization

Table 3 Performance indices of factors induced in multiple aggregations

Cluster	CTF (%)	Purity (%)
Rainfall	100.00	100.00
Temperature	99.91	100.00
Body constitution	92.98	98.29
Checkup	70.65	97.45
Blood analysis	96.99	99.98
Rest	95.67	97.69
General medical	35.49	94.85
Socioeconomic	54.94	94.67
General	13.51	99.57
Lifestyle	69.45	99.91

3.3 Model Validation

The CTF can evaluate the quality of the induced factors. The great advantage of advanced software such as BayesiaLab lies in the possibility for researchers and health professionals to directly calculate this metric-normalized set of values ranging from 0 to 100%. Table 3 shows the CTF values obtained for the latent factors induced in the multi-network design phase.

3.4 Local Impact Analysis of Data

At the fourth level, four supervised Bayesian networks were established, corresponding to each of the defined service activities and whose common node was the worker's state of health. The application of a naïve Bayes algorithm allowed the generation of a pragmatic network structure for the analysis of the influence of each variable on the health status of the workers. 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.

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 important medical conditions affecting these two states are age, total cholesterol, and location, while 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 that there is a 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, cannot be that meaningful from a policy perspective (Table 4).

The great horizontality of variables such as age, location, and total cholesterol directed this study toward the need to add value to those differentiating variables of the musculoskeletal and cardiovascular systems. In addition, according to Table 4, the binary mutual information for the four variables tends to be similar across service activities. This situation leads to the spatial representation of the variable's spine observation, annual precipitation (BIO 12), limb observation, and annual mean temperature (BIO 1) under an ordinary kriging approach (Fig. 5). These variables were selected as the highest contribution to musculoskeletal and cardiovascular systems (Table 4). This approach through OK allows the identification of both a spatial distribution of spinal problems and potentially related extremities, and two differentiated

Table 4 Relative binary mutual information analysis of data over the target node for the states representing musculoskeletal and cardiovascular systems by service activity

System	Variables	Administrative and auxiliary services (%)	Financial and insurance activities (%)	Education (%)	Hostelry (%)
Musculoskeletal	Spine observation	3.47	2.28	3.67	3.23
	Annual rainfall	0.65	0.56	0.75	1.29
Cardiovascular	Limb observation	1.75	1.51	0.88	2.12
	Annual temperature	1.39	0.96	0.45	0.04

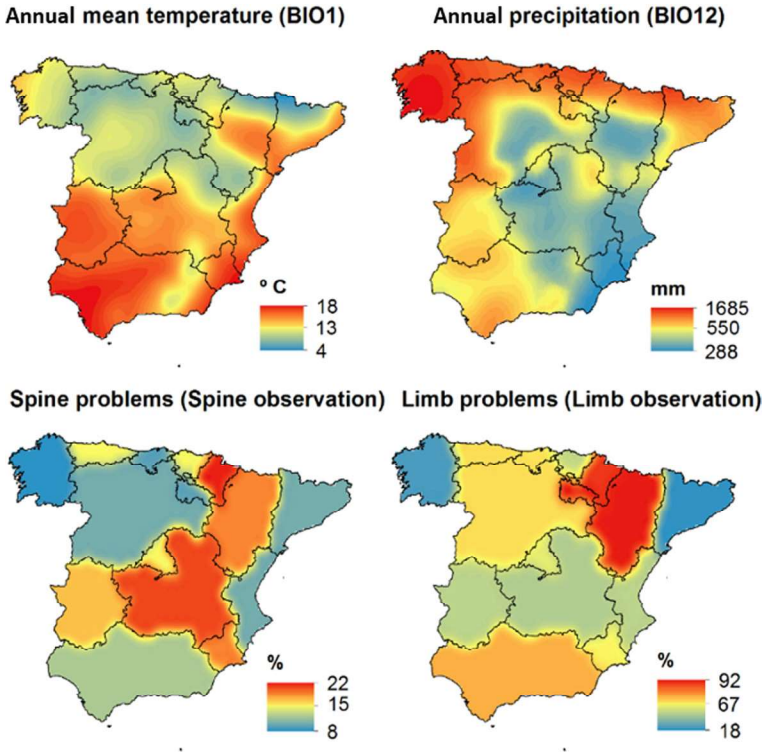

Fig. 5 Distribution maps for annual mean temperature (°C), annual rainfall (mm), and spine and limb observation variables using rate data between 2012 and 2016 interpolated by ordinary kriging

regions where these problems, as well as the systems they are related to, have a higher impact; especially in the northeast of Spain, apart from Catalonia, and the south, with vascular problems such as the presence of varicose veins. As in the western part of Spain, a higher rate of spinal disorders, derived from muscle contraction 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 in-service activities such as hospitality (31.23%) and administration (32.05%). Moreover, we also see that these pathologies are also an underlying cause for problems in the end.

The estimated distribution maps and highly significant clusters were computed by administrative zone for cholesterol, age, and sleep quality (Figs. 6, 7, and 8), giving patterns of high values (red rings) and low values (blue rings). To begin with, as shown in Fig. 6, cholesterol levels will not change with the work area. However, there is a tendency associated with geography. The north is more cholesterol-rich than the south. When the spatial distribution of cholesterol is compared with the distribution of temperatures in Spain, a direct correlation can be identified: the colder the temperature, the higher the cholesterol levels. Previous research indicates that cold increases blood pressure and favors an increase in cholesterol levels (Davis et al. 2022). Another factor

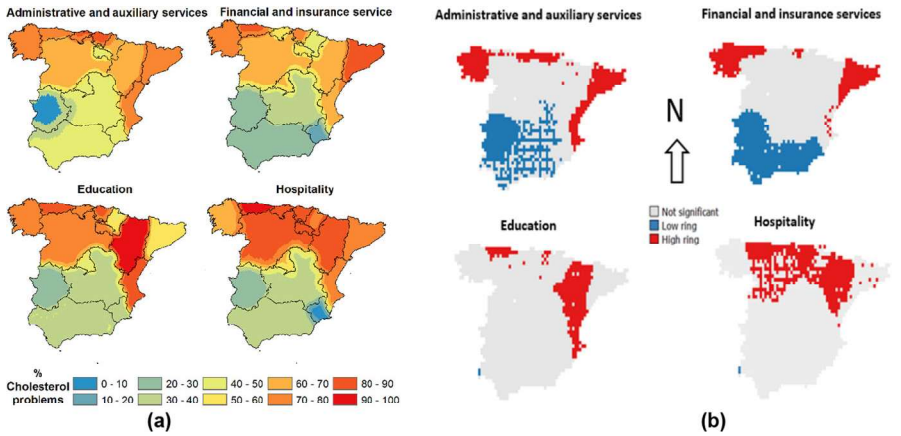

Fig. 6 **a** Distribution maps for cholesterol by service activity group using rate data between 2012 and 2016 through ordinary kriging; **b** high- and low-significance clusters

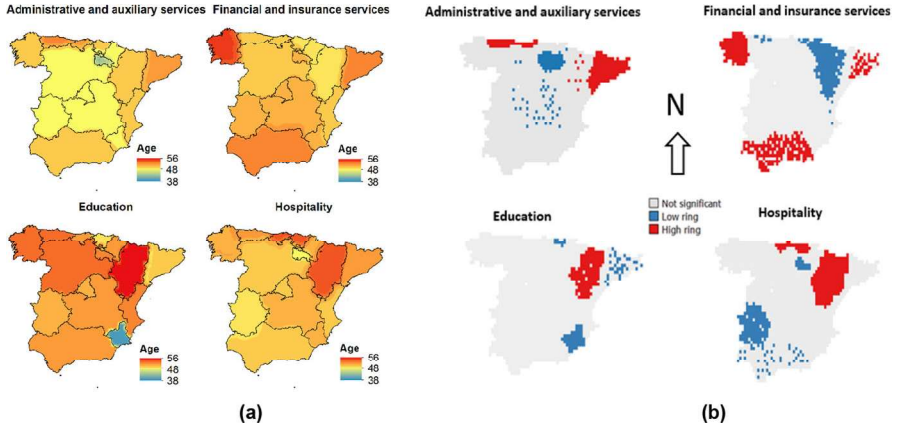

Fig. 7 **a** Distribution maps for age by service activity group using rate data between 2012 and 2016 through ordinary kriging; **b** high- and low-significance clusters

can be the intake of more caloric meals required in northern areas against cold (Tien et al. 2016). However, there are many more factors; the lifestyle of workers and lack of physical activity could also be key factors. In this context, southern regions have a warmer climate that invites more development of physical activities (Bernard et al. 2021).

There is insufficient evidence to establish a correlation between cholesterol (Fig. 6) and age (Fig. 7). In any case, a trend can be observed in which the higher the age, the higher the cholesterol level, a pattern that has been observed in recent studies (Mansoori et al. 2023). Depending on the economic sectors, education has the highest age, except in Murcia, which shows the ageing of the sector in Spain. Looking further into the matter, Aragon has the highest average age and highest cholesterol levels. However,

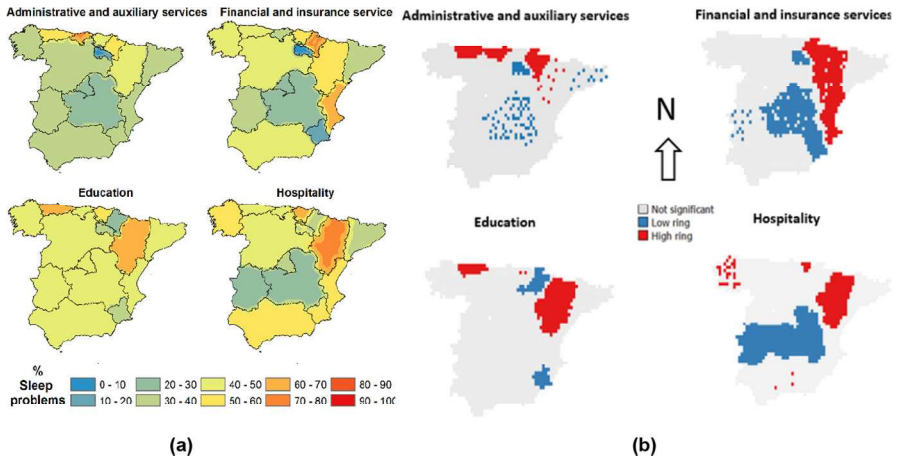

Fig. 8 **a** Distribution maps for sleep quality by service activity group using rate data between 2012 and 2016 through ordinary kriging; **b** high- and low-significance clusters

Catalonia and Navarra have the lowest average ages and the lowest cholesterol levels in the north. In the financial sector, Aragon and Navarre have the lowest average ages and lowest cholesterol levels in the north. However, Galicia and Catalonia have the highest average ages and the highest cholesterol levels in the north (with some exceptions such as Asturias). In the administrative section, La Rioja is the youngest community with the lowest cholesterol levels.

The quality of sleep based on employee service activity is not significantly different (Fig. 8). All data show a similar relationship in terms of color scale variation, except for the education group, which is less blue than the rest. This suggests that the education sector has poorer sleep quality than the rest, except the Navarre region. Freitas et al. (2020) found a similar correlation and linked it to the strong psychological demand teachers have during their psychosocial relationship with students at work. As for correlations with other variables, it is curious that the quality of sleep does not seem to depend upon age. However, sleep disturbances prevail in patients with high diabetes (Huang et al. 2023). This can be corroborated by these data, as regions with the lowest sleep quality can be observed to coincide with regions with high cholesterol levels in the north. Inhabitants of the south of Spain (Andalusia) also present bad sleep quality; in this case, this might be related to high temperatures (Lan et al. 2017).

4 Conclusions

This survey exposes the potentialities of a couple of Bayesian machine learning and geostatistical methodologies in the form of a renewed PSEM to translate the complex problem of determining the occupational health risks of workers in the service sector. The goal is to obtain feasible visual analyses, typical of the geography of health, that

can feed the evolution of medical policies at different levels of the national health system in Spain. This methodology fully applies to the remaining geographic areas. With this methodology, it is important to note that quantifying the influence of bioclimatic and socioeconomic variables becomes a reality. Furthermore, if estimated bioclimatic scenarios are introduced for the coming years, the change in climate attributes may be addressed and consubstantiate insight for future medical decision-making and occupational health.

Concretely, the results of this study revealed that variables such as age, location, and cholesterol, with contributions to the general network between 9 and 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 spine and limb observation, sleep quality, annual precipitation (BIO 12), or annual mean temperature (BIO 1) that, despite not being extremely significant from a mathematical point of view, play a key role and show a great impact in health risk maps at a regional level. It is also worth noting the weight of the bioclimatic variables on the health status of the worker, with a contribution value of approximately 9%. Further analysis is required to measure the uncertainty associated with considering or not considering other groups of variables, including worker behavioral aspects.

Acknowledgements This study was funded by CERNAS-IPCB [UIDB/00681/2020] from the Foundation for Science and Technology (Fundação para a Ciência e Tecnologia—FCT)] and by ICT [UIDB/04683/2020] also from FCT. Carlos Boente obtained a post-doctoral contract within the program PAIDI 2020 (Ref. 707 DOC 01097).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Abad A, Gerassis S, Saavedra Á, Giráldez E, García JF, Taboada J (2019) A Bayesian assessment of occupational health surveillance in workers exposed to silica in the energy and construction industry. Environ Sci Pollut Res 26(29):29560–29569. https://doi.org/10.1007/S11356-018-2962-6/FIGURES/4

Albuquerque MTD, Gerassis S, Sierra C, Taboada J, Martín JE, Antunes IMHR, Gallego JR (2017) Developing a new Bayesian Risk Index for risk evaluation of soil contamination. Sci Total Environ 603–604(2017):167–177. https://doi.org/10.1016/j.scitotenv.2017.06.068

Awotunde JB, Adeniyi AE, Ogundokun RO, Ajamu GJ, Adebayo PO (2021) MIoT-based big data analytics architecture, opportunities, and challenges for enhanced telemedicine systems. Stud Fuzziness Soft Comput 410:199–220. https://doi.org/10.1007/978-3-030-70111-6_10/COVER

BayesiaLab (n.d.) Contingency table fit. Retrieved 22 Dec 2022. https://library.bayesia.com/articles/#!bayesialab-knowledge-hub/key-concepts-contingency-table-fit

Benavoli A, Corani G, Demšar J, Zaffalon M (2017) Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J Mach Learn Res 18(1):2653–2688

Bernard P, Chevance G, Kingsbury C, Baillot A, Romain AJ, Molinier V, Gadais T, Dancause KN (2021) Climate change, physical activity, and sport: a systematic review. Sport Med 51:1041–1059. https://doi.org/10.1007/s40279-021-01439-4

Chang C-H, Shao R, Wang M, Baker NM (2021) Workplace interventions in response to COVID-19: an occupational health psychology perspective. Occup Health Sci 5(1–2):1–23. https://doi.org/10.1007/S41542-021-00080-X/TABLES/1

Conrady S, Jouffe L (2015) Bayesian networks and BayesiaLab: a practical introduction for researchers, vol 9. Bayesia, Franklin

Davis RE, Driskill EK, Novicoff WM (2022) The association between weather and emergency department visitation for diabetes in Roanoke, Virginia. Int J Biometeorol 66(8):1589–1597. https://doi.org/10.1007/S00484-022-02303-4/TABLES/2

Dos Santos LM, Ferraz GAS, Batista ML, Martins FBS, Barbosa BDS (2020) Characterization of noise emitted by a low-profile tractor and its influence on the health of rural workers. An Acad Bras Ciênc 92(3):1–10. https://doi.org/10.1590/0001-3765202020200460

Ebi KL, Vanos J, Baldwin JW, Bell JE, Hondula DM, Errett NA, Hayes K, Reid CE, Saha S, Spector J, Berry P (2021) Extreme weather and climate change: population health and health system implications. Annu Rev Public Health 42:293–315. https://doi.org/10.1146/annurev-publhealth-012420-105026

Eurostat (2022) Contributions of each sector—institutional sector accounts. https://ec.europa.eu/eurostat/web/sector-accounts/detailed-charts/contributions-sectors. Accessed 13 Apr 2022

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. Int J Climatol 37(12):4302–4315

Freitas AMC, de Araújo TM, Fischer FM (2020) Psychosocial aspects at work and the quality of sleep of professors in higher education. Arch Environ Occup Health 75(5):297–306. https://doi.org/10.1080/19338244.2019.1657378

Gao L, Datta A, Banerjee S (2022) Hierarchical multivariate directed acyclic graph autoregressive models for spatial diseases mapping. Stat Med 41(16):3057–3075. https://doi.org/10.1002/SIM.9404

Gerassis S, Abad A, Taboada J, Saavedra Á, Giráldez E (2019) A comparative analysis of health surveillance strategies for administrative video display terminal employees. Biomedl Eng. https://doi.org/10.1186/S12938-019-0737-Z

Gerassis S, Boente C, Albuquerque MTD, Ribeiro MM, Abad A, Taboada J (2021) Mapping occupational health risk factors in the primary sector—a novel supervised machine learning and area-to-point Poisson Kriging approach. Spat Stat 42:100434. https://doi.org/10.1016/J.SPASTA.2020.100434

Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24:189–206

Goovaerts P (1997) Geostatistics for natural resources evaluation. Applied geostatistics series. Oxford University Press, New York, pp 483–837

Ho H, Knudby A, Huang W (2015) A spatial framework to map heat health risks at multiple scales. Int J Environ Res Public Health 12(12):16110–16123. https://doi.org/10.3390/ijerph121215046

Hoffimann J, Augusto J, Resende L, Mathias M, Mazzinghy D, Bianchetti M, Mendes M, Souza T, Andrade V, Domingues T, Silva W, Silva R, Couto D, Fonseca E, Gonçalves K (2022) Modeling geospatial uncertainty of geometallurgical variables with Bayesian models and Hilbert–Kriging. Math Geosci. https://doi.org/10.1007/s11004-022-10013-1

Huang CY, Chen CI, Lu YC, Lin YC, Lu CY (2023) Sleep disturbances, glycaemic control, stress, and coping among diabetic patients: a structural equation modeling approach. Appl Nurs Res. https://doi.org/10.1016/J.APNR.2022.151661

Jerrett M, Gale S, Kontgis C (2010) A companion to health and medical geography. In: Brown T, McLafferty S, Moon G (eds) Wiley-Blackwell, West Sussex, pp 418–445. https://doi.org/10.1002/9781444314762.ch22

Journel AG, Huijbregts CJ (1978) Mining Geostatistics. Academic Press, San Diego

Lahr J, Kooistra L (2010) Environmental risk mapping of pollutants: state of the art and communication aspects. Sci Total Environ 408(18):3899–3907. https://doi.org/10.1016/j.scitotenv.2009.10.045

Keter AK, Lynen L, van Heerden A, Goetghebeur E, Jacobs BKM (2022) Implications of covariate-induced test dependence on the diagnostic accuracy of latent class analysis in pulmonary tuberculosis. J Clin Tuberculosis Mycobacterial Dis 29:10. https://doi.org/10.1016/J.JCTUBE.2022.100331

Kiebish MA, Cullen J, Mishra P, Ali A, Milliman E, Rodrigues LO, Chen EY, Tolstikov V, Zhang L, Panagopoulos K, Shah P, Chen Y, Petrovics G, Rosner IL, Sesterhenn IA, McLeod DG, Granger E, Sarangarajan R, Akmaev V, Dobi A (2020) Multi-omic serum biomarkers for prognosis of disease progression in prostate cancer. J Transl Med 18(1):1–10. https://doi.org/10.1186/S12967-019-02185-Y/TABLES/4

Kirkwood C, Economou T, Pugeault N, Odbert H (2022) Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Math Geosci 54(3):507–531. https://doi.org/10.1007/s11004-021-09988-0

Lan L, Tsuzuki K, Liu YF, Lian ZW (2017) Thermal environment and sleep quality: a review. Energy Build 149:101–113. https://doi.org/10.1016/j.enbuild.2017.05.043

Letta TT, Belay DB, Ali EA (2022) Determining factors associated with cholera disease in Ethiopia using Bayesian hierarchical modeling. BMC Public Health. https://doi.org/10.1186/S12889-022-14153-1

Ley 31/1995, de 8 de noviembre, de prevención de Riesgos Laborales. Boletín Oficial del Estado BOE-A-1995-24292. https://www.boe.es/buscar/pdf/1995/BOE-A-1995-24292-consolidado.pdf

Mansoori A, Sahranavard T, Hosseini ZS, Soflaei SS, Emrani N, Nazar E, Mobarhan MG (2023) Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. Sci Rep 13(1):663. https://doi.org/10.1038/s41598-022-27340-2

Matheron G (1971) The theory of regionalized variables and their applications. Les cahiers du Centre de Morphologie Mathématique, Fascicule 5. Centre de Géostatistique, Fontainebleau, Paris, p 212

Mohamed IN, Mohamed RAF, Hamed A, Elseed M, Patterson V (2021) A children's epilepsy diagnosis aid: development and early validation using a Bayesian approach. Epilepsy Behav. https://doi.org/10.1016/J.YEBEH.2021.108062

Moon G (2020) Health geography. In: International Encyclopedia of human geography, 2nd edn, pp 315–321. https://doi.org/10.1016/B978-0-08-102295-5.10388-9

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press

Njah H, Jamoussi S, Mahdi W (2021) Breaking the curse of dimensionality: hierarchical Bayesian network model for multi-view clustering. Ann Math Artif Intell 89(10–11):1013–1033. https://doi.org/10.1007/s10472-021-09749-z

Orlov A, Sillmann J, Aunan K, Kjellstrom T, Aaheim A (2020) Economic costs of heat-induced reductions in worker productivity due to global warming. Global Environ Change 63:102087. https://doi.org/10.1016/J.GLOENVCHA.2020.102087

Panagos P, Ballabio C, Meusburger K, Spinoni J, Alewell C, Borrelli P (2017) Towards estimates of future rainfall erosivity in Europe based on REDES and WorldClim datasets. J Hydrol 548:251–262. https://doi.org/10.1016/J.JHYDROL.2017.03.006

Pearl J (1998) Graphs, causality, and structural equation models. Sociol Methods Res 27(2):226–284. https://doi.org/10.1177/0049124198027002004

Pearl J (1988) Probabilistic reasoning in intelligent systems, 2nd edn. Morgan Kaufmann, San Francisco, p 552

Perkins-Kirkpatrick SE, Lewis SC (2020) Increasing trends in regional heatwaves. Nat Commun 11(1):1–8. https://doi.org/10.1038/s41467-020-16970-7

Peterson CB, Osborne N, Stingo FC, Bourgeat P, Doecke JD, Vannucci M (2020) Bayesian modeling of multiple structural connectivity networks during the progression of Alzheimer's disease. Biometrics 76(4):1120–1132. https://doi.org/10.1111/BIOM.13235

Rydzik A, Anitha S (2020) Conceptualizing the agency of migrant women workers: resilience, reworking, and resistance. Work Employ Soc 34(5):883–899. https://doi.org/10.1177/0950017019881939

The World Bank Group. (n.d.) Services, value added (% of GDP) | Data. https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS. Accessed 25 Nov 2022

Tien KJ, Yang CY, Weng SF, Liu SY, Hsieh MC, Chou CW (2016) The impact of ambient temperature on HbA1c in Taiwanese type 2 diabetic patients: the most vulnerable subgroup. J Formos Med Assoc 115(5):343–349. https://doi.org/10.1016/j.jfma.2015.03.010

Van Erven T, Harremos P (2014) Rényi divergence and Kullback–Leibler divergence. IEEE Trans Inf Theory 60:3797–3820

Wen C, Huang X, Feng L, Chen L, Hu W, Lai Y, Hao Y (2021) High-resolution age-specific mapping of the two-week illness prevalence rate based on the National Health Services Survey and geostatistical analysis: a case study in Guangdong province, China. Int J Health Geograph 20(1):20. https://doi.org/10.1186/s12942-021-00273-1

Wondmagegn BY, Xiang J, Dear K, Williams S, Hansen A, Pisaniello D, Nitschke M, Nairn J, Scalley B, Xiao A, Jian L, Tong M, Bambrick H, Karnon J, Bi P (2021) 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 773:145656. https://doi.org/10.1016/J.SCITOTENV.2021.145656

World Health Organization (2016) Global strategy on human resources for health: workforce 2030. Switzerland, Geneva

Xu ST, Cao ZC, Huo Y (2020) Antecedents and outcomes of emotional labor in hospitality and tourism: a meta-analysis. Tourism Manag 79:104099. https://doi.org/10.1016/J.TOURMAN.2020.104099

Yousefi L, Tucker A (2022) Identifying latent variables in dynamic bayesian networks with bootstrapping applied to type 2 diabetes complication prediction. Intell Data Anal 26(2):501–524. https://doi.org/10.3233/IDA-205570

Yuvaraj RM, Thulasimala D (2022) Geostatistical analysis of environmental impact on mental health of constructional workers: a case study of Chennai city. In: Hassan MI, Sen Roy S, Chatterjee U, Chakraborty S, Singh U (eds) Social morphology, human welfare, and sustainability. Springer, Cham. https://doi.org/10.1007/978-3-030-96760-4_7