AJAS / SPAS

African Journal of Applied Statistics
Vol. 10 (1), 2023, pages 1368 - 1382.
DOI: http://dx.doi.org/10.16929/ajas/2023.1368.272

ISSN 2316-0861

Quality report of infectious disease modeling techniques for point-referenced spatial data: A Systematic review

Romuald Beh Mba[(1,*)], **Bruno Enagnon Lokonon**[(1)] **and Romain Glèlè Kakaï**[(1)]

[(1)] Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF), Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, 04 BP 1525, Cotonou, Bénin

Received on June 23, 2023; Accepted on August 04, 2023; Published on October 19, 2023

Abstract. Spatial data modeling can provide significant value to healthcare organizations by improving decision support, resource management and distribution, and clinical outcomes. The aim of this study was to (i) summarize the trends of the modeling techniques used to analyze point-referenced spatial data in epidemiology and (ii) examine if all information required when applying these modeling techniques were properly reported in the published papers. A literature search was limited to journal papers published from January 2010 to June 2022 using PubMed, Scopus, Crossref, and Google Scholar. From 528 articles identified with the defined keywords, 351 were retained for the review [...](See Complete sbstract in page 1369)

Key words: spatial models; disease mapping; smoothing methods; quality assessment; epidemiology

AMS 2010 Mathematics Subject Classification Objects : 62H11; 60G60; 62F15.

[(*)] Corresponding author: Romuald Beh Mba (behmbaroms@gmail.com)
Bruno Enagnon Lokonon: brunolokonon@gmail.com
Romain Glèlè Kakaï: glele.romain@gmail.com

Full Abstract in English. Spatial data modeling can provide significant value to healthcare organizations by improving decision support, resource management and distribution, and clinical outcomes. The aim of this study was to (i) summarize the trends of the modeling techniques used to analyze point-referenced spatial data in epidemiology and (ii) examine if all information required when applying these modeling techniques were properly reported in the published papers. A literature search was limited to journal papers published from January 2010 to June 2022 using PubMed, Scopus, Crossref, and Google Scholar. From 528 articles identified with the defined keywords, 351 were retained for the review. The results revealed that the use of modeling techniques in spatial data for infectious diseases increases exponentially over time. The most common spatial method was Empirical Bayesian Kriging [EBK] (52% of the selected articles), followed by Spatial GLMMs (34%) and Spatial smoothing Kernel Estimation (13%).

Résumé (Abstract in French) La modélisation des données spatiales peut apporter une valeur significative aux organisations de santé pour la prise de décision en matière de recherche clinique. L'objectif de cette étude avait pour but (i) d'étudier les tendances des techniques de modélisation utilisées pour analyser les données spatiales ponctuelles en épidémiologie et (ii) de voir si toutes les informations requises lors de l'application de ces techniques étaient correctement rapportées dans les articles publiés. Une recherche documentaire a été limitée aux articles publiés entre janvier 2010 et juin 2022 en utilisant PubMed, Scopus, Crossref et Google Scholar. Sur les 528 articles identifiés à l'aide des mots-clés définis, 351 ont été retenus. Les résultats ont révélé que l'utilisation de ces techniques de données spatiales augmente de façon exponentielle. La méthode spatiale la plus courante est le krigeage bayésien empirique [EBK] 52%, suivi par le GLMM spatial (34%) et l'estimation par noyau de lissage spatial (13%)

Presentation of all authors

Romuald BEH MBA, M.Sc, is preparing his PhD thesis in Biometry, at University of Abomey-Calavi (UAC) and at Laboratoire de Biomathématiques et d'Estimation Forestières (LABEF), Benin, under the supervision of ...

Bruno Enagnon Lokonon, Ph.D., Research fellow in Biometry and Ecology at: Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF), University of Abomey-Calavi (UAC), Benin

Romain Glèlè Kakaï, Ph.D., is a full professor of Professor of Biometrics and Forest estimations at: University of Abomey-Calavi (UAC) Benin. He is the Head of the Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF), the Coordinator of the doctoral studies in Biometry and the President of the Scientific Council of Agronomic Sciences.

1. Introduction

modeling point-referenced spatial data can provide significant value to healthcare organizations by improving decision support and policy recommendations (Cramb et al., 2018). Spatial epidemiology studies the spatial distribution of disease incidence and its relationship to the potential risk factors. Its origins go back to 1855 with the seminal work of Snow on cholera transmission, where he mapped the cholera cases with the locations of water sources in London and showed that contaminated water was the major cause of the disease (Shiode et al., 2015). Spatial epidemiological tools applied in infectious disease research can identify areas of high disease transmission and assess the potential environmental and other risk factors that can explain space variation (Chowell and Rothenberg, 2018). There are many potential methodological approaches to examining spatial infectious disease data.

Point-referenced spatial data arise from observations collected at geographical locations over a fixed continuous space. Proximity in space introduces correlations between observations invalidating the independence assumption of standard statistical methods. Ignoring spatial correlation will result in underestimation of the standard error of the parameter estimates (Poggio et al., 2016). Using spatial models to model infectious diseases can be complex and challenging, particularly since numerous factors, including environmental, demographic, and socioeconomic factors, may influence the spread of infectious diseases. Several potential problems can arise when using spatial models in modeling infectious diseases. One problem is the issue of data quality and availability. Spatial modeling requires accurate and high-quality data, which may not always be available, particularly in resource-limited settings. This can lead to biased or incomplete models, which may not accurately reflect the actual spread of the disease (Jones et al., 2009). Another problem is the complexity of the models themselves for the users. Spatial models often involve complex mathematical and statistical techniques that can be difficult to understand and interpret. Moreover, different modeling approaches may produce different results, making it difficult to determine the most appropriate approach. A systematic literature review can help identify the strengths and weaknesses of different modeling approaches and provide guidance on which approach is most appropriate for a given research question or dataset (Malhotra, 2015). In addition, spatial models are often used to inform public health policies and interventions. However, the effectiveness of these policies and interventions may be influenced by factors such as social and cultural norms, political will, and healthcare infrastructure (Williams et al., 2015). Overall, a systematic review of the literature is important in addressing the potential problems associated with the use of spatial models in infectious diseases modeling. By identifying sources of bias, suggesting appropriate modeling approaches, and considering contextual factors, such reviews can improve the accuracy and usefulness of spatial models in informing public health policies and interventions (Barros et al., 2020).

Improving decision support, resource management and distribution, and clinical outcomes are valuable and critical tasks for healthcare organizations. This may explain the increasing adoption of spatial data and space-based computing techniques. However, there is still a scarce understanding of the estimation methods used for point-referenced spatial data in infectious diseases modeling (Cowan, 2013). A systematic review of Point referenced spatial data modeling techniques has yet to be conducted to foster the correct application of these approaches. In this regard, this study aims to systematically review the literature on the model used for point-referenced spatial in modeling infectious diseases. Specifically, it aims to: **(1)** summarize the trends of the modeling techniques used to analyze point-referenced spatial data in epidemiology; **(2)** examine if all results from the application of these modeling techniques are properly reported in the published papers; **(3)** discuss the advantages and disadvantages of the modeling techniques used to analyse point-referenced spatial data; **(4)**propose research perspectives.

2. Material and methods

2.1. Search Strategy

The paper search was carried out to include only studies in Epidemiology that were published in English. The topics and acronyms considered are all appertained to GLMMs (i.e., generalized mixed model, hierarchical generalized model, multilevel generalized model, GLMM, and HGLM).In a first step, we identify and describe all models that allow estimating parameters for point-referenced spatial data in GLMMs. Biomedical databases (PubMed) and science databases (Scopus, Crossref, and Google Scholar) were searched electronically. The literature search was limited to peer-reviewed journal papers published from January 2010 to June 2022. Papers were searched based on the same keywords, “generali*linear mixed” OR “hierarchical generali*linear” OR HGLM* OR “multilevel generali*linear” OR GLMM* OR spatial models OR Bayesian models OR disease mapping. A Boolean driver was enforced to link the keywords. All results were combined, and the duplicates were removed using EndNote. The titles and abstracts of papers set up through keyword quests were screened first, and full texts were read when the papers were relevant. This review was conducted based on the Preferred Items for Systematic Reviews and Meta-analyses (PRISMA) statement (Moher et al., 2010).(See Figure 1, page 1372).

2.2. Selection Criteria

Firstly, abstracts of all records retrieved from January 2010 to June 2022 were screened, excluding duplicates, conference papers, book reviews, theoretical studies (statistical tests, new procedures, mathematical developments, comparison of

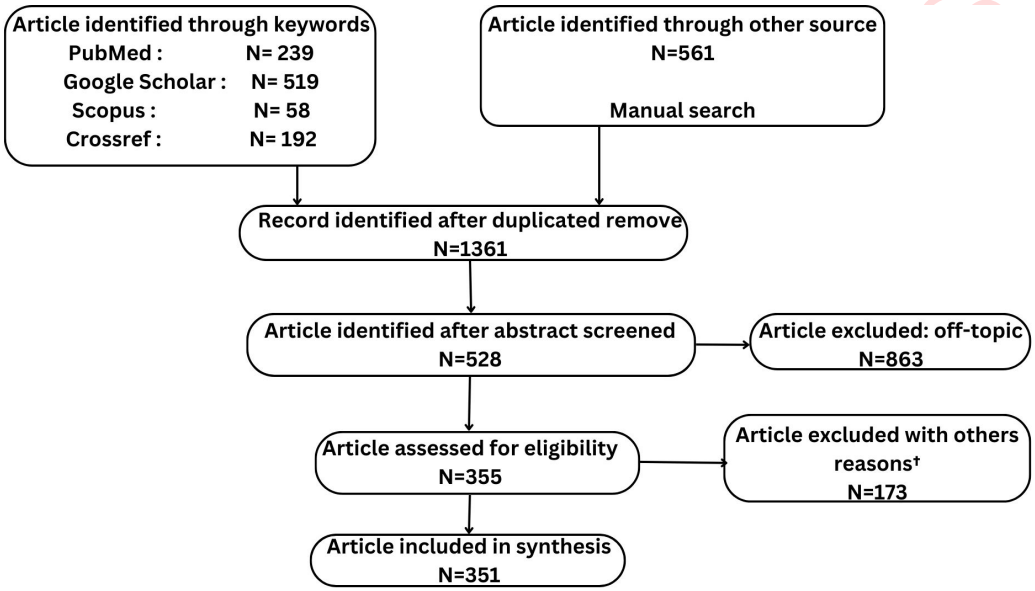

Fig. 1: Flowchart showing the selection of articles published in the last twelve (12) years until June 2022 and included in the systematic review

models, etc.), illustrations or tutorials of the utilization of analytic techniques, testing models of study, simulation studies, description of applied math software package, systematic reviews, and empirical studies in fields unrelated to medical speciality and infectious disease. Of the 528 papers originally retrieved, 173 were excluded. The inclusion criteria were as follows. Studies were enclosed if they used modeling techniques or Spatiotemporal theorem models. Papers were excluded if they solely projected applying GLMMs (e.g., study protocols) or contained inconsistencies or issues within the cryptography of variables

2.3. Quality assessment

To assess the quality of spatial methods in modeling infectious diseases, several factors or characteristics have been evaluated, such as: **(i)** Data quality: The quality of the data used in the study, including the distribution, the accuracy and reliability, and the methods used to collect and analyze the data. **(ii)** Model specification: the specification of the spatial methods, including the choice of the response variable, the fixed and random effects, and the distributional assumptions. **(iii)** Model fit: we evaluated the goodness-of-fit statistics, including residual plots and model diagnostics. **(iv)** Model validation: using appropriate methods, such as cross-validation or bootstrapping, to ensure that the model

do not overfit the data and the **(v)** Model interpretation: The interpretation of the results includes the estimation of parameters, the inference of statistical significance, and the interpretation of the model coefficients.

2.4. Data extraction and statistical analyses

To analyze the trend of the use of modeling techniques between January 2010 and June 2022, we recorded the year of publication, journal title, and country of affiliation of the primary author. Information also included the topics addressed within the study and the modeling technique used. We accustomed count and relative frequencies to explain their trend. All analyses were done within R software (version 4.1.2).

3. Results

3.1. Overview of the modeling techniques used for spatial public health data

The evolution of the use of modeling techniques in the 528 selected papers reveals that spatial data for modeling infectious diseases increases over time (Figure 2, page 1374). In the final step, we selected 351 papers to evaluate the trend and the quality of report information. Results give the summary frequency of spatial methods and substantive focus; graphs explored trends over time. The most common spatial methods were Empirical Bayesian kriging (52%), followed by spatial Generalized Linear Mixed Models (34%), spatial smoothing Kernel Estimation (13%), and Artificial Neural Network Model (1%) (Table 1, page 1374).

This table shows that the Empirical Bayesian kriging is still the most common Spatial method published in Springer: Nature (68.63%), followed by Biomedical journals (56.63%) and BMC infectious disease (55.26%). Spatial Generalized Linear Mixed Models represent (44% and 36.84%) of articles published in Spatial Statistics and BMC infectious disease journal, respectively. (Table 1, page 1374).

3.2. Quality assessment of the modeling techniques for spatial public health data

The characteristics used to evaluate the quality of the information reported in the reviewed papers are presented in Table 2, page 1375. It results that the most popular distribution used is the Poisson (34.69%), followed by the Normal (28.57%) and the Binomial (18.37%). In total, 18.37% of the papers evaluated did not report the distribution used. In the selected papers, 12.24% and 71.43% specified the fixed and the random effect test, respectively, 20.41% specified the Goodness of fit, and 20.41% specified the Residual plot. Most of the papers did not specify the model fit.

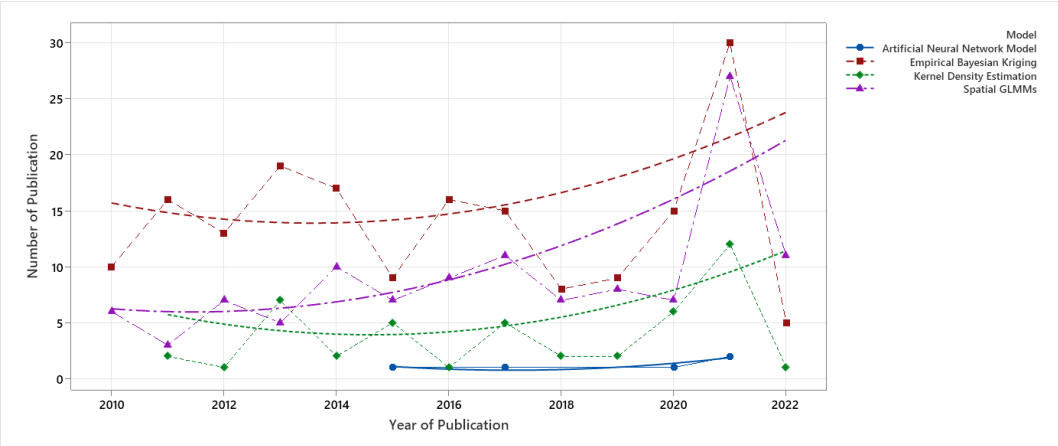

Fig. 2: Number of articles by year of publications

Table 1: Summary of the structure of the spatio-temporal models discussed in the selected papers.

Journals	SGLMMs	EBK	ANN	KDE	**TOTAL**
	N (%)	N (%)	N (%)	N (%)	N (%)
Biometrical Journal	14 (23.72)	34 (56.63)	3 (5.08)	8(13.56)	**59 (16.81)**
Infectious disease	28 (36.84)	42 (55.26)	1 (1.32)	5 (6.58)	**76 (21.65)**
Spatial Statistics	44 (44)	40 (40)		16 (16)	**100 (28.49)**
PlosONE	21 (32.31)	31 (47.69)	1 (1.54)	12 (18.46)	**65 (18.52)**
Springer: Nature	11 (21.57)	35 (68.63)		5 (9.80)	**51 (14.53)**
TOTAL	118 (34)	182 (52)	5 (1)	46 (13)	**351 (100)**

Legends: SGLMMS=Spatial GLMMS; EBK=Empirical Bayesian Kriging; ANN= Artificial Neural Network; KDE= Kernel Density Estimation

The cross-validation was reported in (4.08%) and Bootstrapping in (10.2%). Statistical modeling (53.06%), Inference (38.78%) and the regression coefficient (42.86%) were reported. Overdispersion was specified in (38.18%) of the review articles. However, many papers did not report all the requested parameters. This represents a very high percentage of analyses (82.2%), and other estimation methods may have also been used. Overall, assessing the quality of modeling techniques for spatial public health data requires a thorough understanding of the data, the modeling approach, and the interpretation of results. It is important to choose appropriate methods, validate the model, and communicate the results meaningfully and transparently to stakeholders.

Table 2: The quality evaluation of some parameters in the review article.

Model Specification			**Model Fit**			**Model validation**			**Model Interpretation**		
Distribution	N	(%)	Goodness -of-fit	N	(%)	cross validation	N	(%)	Statistical modeling	N	(%)
Binomial	63	18.37	Specified	70	20.41	Specified	14	4.08	Specified	182	53.06
Normal	98	28.57	Not Specified	273	79.59	Not Specified	329	95.9	Not Specified	161	46.94
Poisson	119	34.69									
Not Specified	63	18.37	Residual plots			Bootstrapping			Inference		
Fixed Effects Test			Specified	70	20.41	Specified	35	10.2	Specified	133	38.78
Specified	42	12.24	Not Specified	273	79.59	Not Specified	308	89.8	Not Specified	210	61.22
Not Specified	301	87.76									
Random Effects Test						Overdispersion Evaluation			Coefficients		
Specified	245	71.43				Yes	131	38.2	Specified	147	42.86
Not Specified	98	28.57				No	212	61.8	Not Specified	196	57.14

3.3. The capabilities of different software packages and functions for SGLMM analysis

A growing number of software-related packages focus on theoretical spatial models. Within these various classes of models, some different packages and functions are used in R, Python, and other software (Table 3, page 1376). A multitude of packages allows for spatial analysis in **R**, including methods appropriate for point and area-level data, such as: gdistance: computes distances and routes on geographic grids; *geoR:*for inference in generalized linear spatial models using Markov chain Monte Carlo (MCMC); *geospacom:* generates distance matrices from shapefiles and plots the data on maps;*spatsurv:* Bayesian inference for parametric spatial survival models with proportional hazards; *spdep:* useful functions for creating spatial weight matrix objects from polygon adjacencies, and various global and spatial correlation tests; *sphet*: Estimation of spatial autoregressive models with and without heteroskedastic innovations. *PrevMap* performs spatial prediction, setting model parameters to the most random Monte Carlo estimates of a binomial geostatistical supply model. **Phyton** with the functionality PySAL or PyMC3 can be a new open-source PP framework with an intuitive and readable interface that performs Bayesian statistical modeling and model fitting centred on advanced Markov chain Monte Carlo and variational fitting algorithms.

3.4. Decision tree for choosing spatiotemporal epidemiological tools

Here, we are presenting a decision tree for choosing Spatiotemporal visualization and Analytical Tools (SATs) in Figure 3, page 1377. This decision tree is assessed into three stages: **(A)** pre-hypothesis: testing/hypothesis-generating stage that refers to the existence of spatial dependence and spatial patterns within the distribution of adverse health events; **(B)** primary hypothesis testing stage that involves the association of the events with risk factors/covariates and **(C)** secondary-hypothesis testing and spatial modeling stage wherever the predictions and inferences are made. The various SATs are broadly classified into four categories: visualization and descriptive analysis; spatial/Spatiotemporal dependence and pattern recognition; spatial smoothing and interpolation; and spatial correlation studies:

Table 3: Common mapping, GIS and statistical software and capabilities used in the 351 SGLMM analyses

						Smoothing		Model-based smoothing			
Software	Type	Visualise maps	Neighbour-hood matrix	Spatial correlation	Raw estimates	Locally-weighted	Kernel	Spatial regression	Poisson kriging	EB	HB
Open source											
Bing Maps	Map	Y									
BUGS	Stat	Y	Y	Y	Y						
Epi Info	Tools	Y			Y						
GeoDa	Tools	Y	Y	Y	Y	Y	Y	Y		Y	
GRASS	GIS	Y									
Google Earth	Map	Y									
JAGS	Stat										Y
NIMBLE	Stat										Y
PySAL	Tools	Y	Y	Y	Y	Y	Y	Y		Y	Y
R	Stat	Y	Y	Y	Y	Y	Y	Y		Y	Y
SaTScan	Tools	Y		Y				Y			
Stan	Stat										Y⋆
Commercial											
ArcGIS	GIS	Y	Y		Y	Y	Y	Y			
MapInfo	GIS	Y	Y		Y	Y	Y				
MLwiN	Stat							Y			Y⋆
SAS	Stat	Y	Y	Y	Y	Y	Y⋆	Y		Y	Y
S-Plus	Stat	Y	Y	Y	Y						
SpaceStat	Tools	Y	Y	Y					Y		
Stata	Stat	Y	Y	Y	Y	Y	Y	Y		Y	Y
TerrSet	GIS	Y	Y	Y	Y	Y	Y				

Abbreviations: Stat=statistical software, Map=mapping software, GIS=geographic information systems software, EB=Empirical Bays, HB=Hierarchical Bays, Y=Yes.
⋆ Indicates limited functionality, such as no pre-programmed CAR distributions. Note that software can often interconnect with others to provide greater functionality (e.g., statistical packages and GIS software) or to facilitate programming in the appropriate language (e.g., Stan and JAGS interface with R). Softwares can perform a hierarchical Bayes analysis if a random effects term for each area can be modeled.

modeling and regression. The decision tree seeks to recommend an acceptable class of the SATs supported by the stage of the research question. SATs are usually utilized in medical speciality studies. It is vital to notice, however, that this is often not a scientific review of the Spatiotemporal visualization and Analytical Tools, which the classification used here is somewhat arbitrary, given the subjective nature of the matter. This contribution of a systematic review, whereas not associated thoroughgoing description of SATs, intends to produce a brief guide to introductory-level population and ecological scientists on ordinarily used tools and encourage the users to explore the varied algorithms for a lot of familiar conclusions. Elaborated reviews on SATs will be found elsewhere, as well as a gloss of ordinarily used terms and their definitions in spatial medicine.

4. Discussion

4.1. State of usage of Spatial Methods for modeling infectious disease

This study focused on determining the extent to which information or data are spatially auto-correlated and performing hypothesis tests after accounting for spatial auto-correlation (F. Dormann et al., 2007). Epidemiological models have become more important for studying the spread of diseases, designing interven-

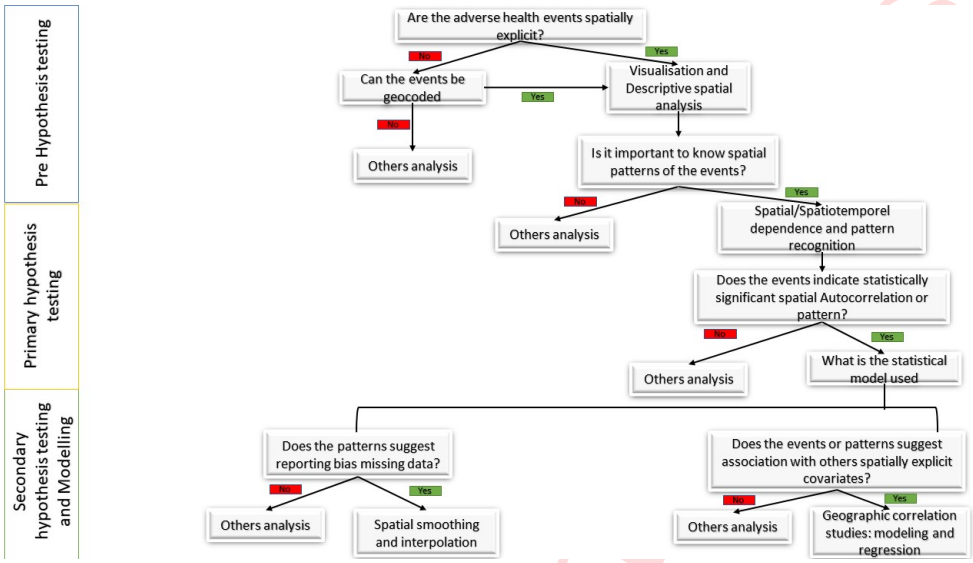

Fig. 3: A simplistic illustration of a framework for choosing Spatiotemporal visualization and Analytical Tools (SATs)

tions to monitor and prevent new epidemics, and reducing their devastating effects on a population. This study aimed to systematically review literature from January 2010 to June 2022 on the model used for point-referenced spatial by presenting the methodology of the models used in the epidemiology study, often used to understand and predict diseases (infectious and non-infectious) occurring in a given region. We also provide more specific information about how these models are performed and reported in point-referenced spatial data. This enabled us to identify gaps in the presentation of results and, therefore, to evaluate the quality of reports involving GLMMs in point-referenced spatial data. The number of papers using the spatial models in different areas of point-referenced spatial data increased over the period 2010 – middle of 2022, although at a slower rate than was observed in clinical drugs by Casals et al. (2014) between 2000 and 2012. This is analogous to what passed with LMMs, employed previously to GLMMs, which were applied first in drugs and more gradationally in psychology (De Bono et al., 2008). Thus, our stopgap is that GLMMs will also be extensively used in point-referenced data once researchers in this field become more apprehensive of their advantages. Our review of the period 2010 – middle 2022 included papers linked to the area of disease mapping from JCR- listed journals. The journal with the largest number of papers involving Spatial GLMMs was Spatial Statistics, from which 100 of the retrieved articles satisfied the inclusion criteria for this review. It, however, provided 18 publications in only 2021 but was the highest number of publications among all the years per all five journals. Nevertheless, only one publication

was reviewed in 2010 for this same journal. Regarding the characteristics used to evaluate the quality of the information reported in the review articles, the Poisson distribution was the most popular, followed by the Normal and Binomial distributions. More than half of the GLMM analyses we reviewed did not report the shape of the distribution or the link function. Although there is a natural association between these two variables, this information must be returned as different link functions that may be suitable for a given distribution. Furthermore, in the event of overdispersion specified in the review articles, it is usual to use an alternative distribution to the one fitting the data, for example, Poisson data, which frequently presents overdispersion. Regarding statistical software, we did not quantify the strategies for model building. The review presents some of the main packages in the statistical software and helps implement the mentioned spatial models.

4.2. Advantages and disadvantages of the most Spatial Methods in the reviewed paper

According to our review article, there are more smoothing techniques, and they can be categorized as global and local smoothing techniques. Kernel smoothing, one of the widely used techniques, facilitates visualization of the intensity of events while accounting for background spatial distribution of the population at risk (Diggle and Giorgi, 2016) and generates tolerance contours for which the relative risk of disease is significantly high (Kanankege et al., 2020). The kernel smoothing method describes and visualizes health threats' intensity or spatial relative risk. So smoothing techniques are used to reduce noise by shrinking values toward the adjacent observations and estimate the spatial trend, which applies to homogeneous and heterogeneous point processes (Kanankege et al., 2020). In the heterogeneous point process in which the intensity of the spatially varying event change within the study area, smoothing is used to increase the accuracy of estimating the event intensity using either parametric or non-parametric methods (Kanankege et al., 2020). Spatial smoothing techniques utilize a moving weighted function to reduce noise by emphasizing the differences between values on a surface, producing a spatially continuous map. Although KDE is a powerful and flexible technique, it has some limitations, such as:

(i) Choosing the bandwidth parameter: The choice of the bandwidth parameter significantly impacts the performance of KDE. If the bandwidth is too small, the estimate will be highly variable, while if the bandwidth is too large, the estimate will be over-smoothed and may not capture the underlying distribution accurately. Choosing an appropriate bandwidth can be difficult and may require some trial and error.

(ii) Sensitivity to outliers: KDE is sensitive to outliers in the data. Outliers can significantly affect the density estimate, and a small number of outliers can lead to a biased estimate.

(iii) Curse of dimensionality: KDE becomes computationally intensive as the number of dimensions increases. As the number of dimensions increases, the number of observations needed to obtain an accurate estimate grows exponentially. This makes KDE impractical for high-dimensional data sets.

(iv) Boundary effects: KDE can be affected by boundary effects, especially if the data is not well distributed over the domain. In such cases, the density estimate may not accurately capture the true density near the boundaries. (v) Interpretation: KDE provides an estimate of the probability density function, which can be difficult to interpret for non-experts. The estimate provides no information about the underlying distribution, such as moments or quantiles. Despite these limitations, KDE remains a useful and widely used method for density estimation in various fields, including statistics, economics, and finance.(Kanankege et al., 2020) and headbanging considered as alternatives for detecting circumscribing clusters of varying shapes instead of circular clusters.

Empirical Bayesian kriging (EBK) is a spatial interpolation method used in Geostatistics to estimate the value of a variable at unsampled locations. It is a hybrid approach that combines the strengths of two different techniques: kriging and empirical Bayesian methods. EBK has several advantages over other interpolation methods. First, it allows for the incorporation of prior knowledge about the spatial correlation structure of the data into the interpolation process. Second, it can be used to generate probabilistic estimates of the interpolated values, which can be useful in many applications. Finally, EBK is computationally efficient and can handle large datasets.(Gribov and Krivoruchko, 2020). If two countries have the same standardized incidence ratio (SIR), for example, but have different population sizes, the confidence of EBS estimates would be higher in the county with larger population size. The key feature of using EBK for count data interpolation and regression analysis is the appropriate data transformation to the Gaussian process and direct approximation of the discrete data distribution by one Gaussian distribution, which has been suggested and discussed by Singh and Oliveira (Singh et al., 2019) and (Oliveira et al., 2018). However, EBK does have some limitations. It assumes that the spatial correlation structure of the data is stationary, meaning that it does not vary over space. It also assumes that the data are normally distributed, which may not always be true in practice. Additionally, EBK requires some expertise in Geostatistics and statistical modeling to implement properly.

Spatial Generalized Linear Mixed Models (GLMMs) are statistical models that allow for the inclusion of spatial autocorrelation in the response variable. It plays a

significant role in modeling spatial and spatiotemporal patterns of infectious diseases. Spatial GLMMs can account for spatial autocorrelation, the tendency of observations close together in space to be more similar than observations far apart. By accounting for spatial autocorrelation, the model can provide more accurate estimates of the relationship between the response and predictor variables. Spatial GLMMs can improve the accuracy of predictions for new locations, especially when the spatial autocorrelation is strong. This is because the model can borrow strength from neighbouring locations to make more accurate predictions. Spatial GLMMs can account for unobserved heterogeneity not captured by the predictor variables. This can lead to more accurate estimates of the parameters of interest and better predictions. (Lowe et al., 2011). Two other studies used an appropriate CAR priority [(Hu et al., 2011), (Samat and Percy, 2012)]. Spatial Generalized Linear Mixed Models (GLMMs) also present disadvantages such as (i) Complexity: Spatial GLMMs are more complex than traditional GLMMs because they involve modeling the spatial autocorrelation structure. This can make the models harder to fit and interpret; (ii) Requires large sample size: Spatial GLMMs require a large sample size to estimate the parameters accurately. This is because the spatial autocorrelation structure needs to be estimated from the data. This requires a sufficient number of observations; (iii) Computationally intensive: Fitting Spatial GLMMs can be computationally intensive, especially when the number of observations is large, or the spatial autocorrelation structure is complex. This can make the models computationally expensive to fit and may require specialized software or hardware. However, specific areas, such as rural or neighborless areas, need to be studied to improve the correlation structure in the model. For example, distance-based weight matrices may be preferable for studying the effect of road travel or human mobility. Only one study used different types of neighborhood adjacency matrices, namely binary, weighted by border length and weighted by border and barriers (Ferreira and Schmidt, 2006).

5. Conclusion

This study conducted a systematic review of the literature on spatial models. it examined published articles to identify and evaluate the quality of reporting on spatial models in the field of disease mapping-based epidemiology. After a careful review of the content of these articles, we can conclusively state that the use of spatial models in the epidemiology literature has increased over the years. Based on this systematic literature review, we conclude that spatial model techniques have been widely used in modeling infectious diseases and have proven useful for exploring the spatial and temporal patterns of disease spread. However, it is important to ensure that the models are specified and fitted correctly and that the results are interpreted appropriately to obtain reliable and accurate results. The objectives of the analysis, the quality of the data, and the expected outcomes can all influence the choice of the final model. Nevertheless, the Empirical Bayesian Kriging method is increasingly used for disease mapping and has been

shown to perform well overall. With the more recent application of approximation methods, we are able to generate results quickly. This review excluded studies that did not perform a defined spatial analysis, such as calculating a spatial metric, examining local relationships between spatially contiguous entities, or using a spatial statistic, even if the study involved spatial exposures or outcomes. Even among the reviewed papers, spatial interpolation or smoothing methods for estimating geographically variable exposures and using spatial regression alone or in combination with multilevel models have been underutilized. These techniques should be more widely applied because they can improve the specificity of exposure-disease relationships, reduce measurement error, and deepen our understanding of place-health relationships. Therefore, it is important to consider using minimal rules as guidelines when using spatial methods.

Acknowledgment. This work was financially supported by the German Academic Exchange Service, or DAAD. We thank this institution and its donors.

We are extremely grateful to the editor and the anonymous reviewer for their valuable comments and suggestions, which have Helped to improve the quality of our manuscript.

References

Barros, J. M., Duggan, J., and Rebholz-Schuhmann, D. (2020). The application of internet-based sources for public health surveillance (infoveillance): systematic review. *Journal of medical internet research*, 22(3):e13680.

Chowell, G. and Rothenberg, R. (2018). Spatial infectious disease epidemiology: on the cusp.

Cowan, N. M. (2013). *A geospatial data management framework for humanitarian response*. PhD thesis, The George Washington University.

Cramb, S., Duncan, E., Baade, P., and Mengersen, K. (2018). Investigation of bayesian spatial models.

De Bono, J. S., Scher, H. I., Montgomery, R. B., Parker, C., Miller, M. C., Tissing, H., Doyle, G. V., Terstappen, L. W., Pienta, K. J., and Raghavan, D. (2008). Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. *Clinical cancer research*, 14(19):6302–6309.

Diggle, P. J. and Giorgi, E. (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *Journal of the American Statistical Association*, 111(515):1096–1120.

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.

Ferreira, G. S. and Schmidt, A. M. (2006). Spatial modelling of the relative risk of dengue fever in rio de janeiro for the epidemic period between 2001 and 2002. *Brazilian journal of Probability and Statistics*, pages 29–47.

Gribov, A. and Krivoruchko, K. (2020). Empirical bayesian kriging implementation and usage. *Science of the Total Environment*, 722:137290.

Hu, W., Clements, A., Williams, G., and Tong, S. (2011). Spatial analysis of notified dengue fever infections. *Epidemiology & Infection*, 139(3):391–399.

Jones, D. A., Wang, W., and Fawcett, R. (2009). High-quality spatial climate data-sets for australia. *Australian Meteorological and Oceanographic Journal*, 58(4):233.

Kanankege, K. S., Alvarez, J., Zhang, L., and Perez, A. M. (2020). An introductory framework for choosing spatiotemporal analytical tools in population-level eco-epidemiological research. *Frontiers in Veterinary Science*, 7:339.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A., Carvalho, M. S., and Barcellos, C. (2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in brazil. *Computers & Geosciences*, 37(3):371–381.

Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing*, 27:504–518.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., et al. (2010). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery (London, England)*, 8(5):336–341.

Oliveira, C. B., Maher, C. G., Pinto, R. Z., Traeger, A. C., Lin, C.-W. C., Chenot, J.-F., van Tulder, M., and Koes, B. W. (2018). Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *European Spine Journal*, 27(11):2791–2803.

Poggio, L., Gimona, A., Spezia, L., and Brewer, M. J. (2016). Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in scotland using r-inla. *Geoderma*, 277:69–82.

Samat, N. and Percy, D. (2012). Vector-borne infectious disease mapping with stochastic difference equations: an analysis of dengue disease in malaysia. *Journal of Applied Statistics*, 39(9):2029–2046.

Shiode, N., Shiode, S., Rod-Thatcher, E., Rana, S., and Vinten-Johansen, P. (2015). The mortality rates and the space-time patterns of john snow's cholera epidemic map. *International journal of health geographics*, 14(1):1–15.

Singh, D., Agusti, A., Anzueto, A., Barnes, P. J., Bourbeau, J., Celli, B. R., Criner, G. J., Frith, P., Halpin, D. M., Han, M., et al. (2019). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the gold science committee report 2019. *European Respiratory Journal*, 53(5).

Williams, E. P., Mesidor, M., Winters, K., Dubbert, P. M., and Wyatt, S. B. (2015). Overweight and obesity: prevalence, consequences, and causes of a growing public health problem. *Current obesity reports*, 4:363–370.