WILEY

AREA

Trends and topics in Geographically Weighted Regression research from 1996 to 2019

Journal:	*Area*
Manuscript ID	AREA-RP-Nov-2020-0133
Manuscript Type:	Regular Paper
Keywords:	geographically weighted regression, bibliometric analysis, topic modelling, latent Dirichlet allocation, collaboration pattern, machine learning

Abstract:	This research was conducted in order to improve the understanding of the structure, contents, and trend of topics within the existing literature in the field of geographically weighted regression. Additionally, it is intended to determine and produce a mapping of scientific networks in the domain of geographically weighted regression. The proposed methodology implements a combination of bibliometric techniques and modelling of topics in order to extract the latent topics from the collected literature by utilizing latent Dirichlet allocation and machine learning tool. The results achieved allowed to stress the most prolific authors, the most cited authors, the most representative articles and journals and the countries which are responsible for the publications.

1. INTRODUCTION

Geographically Weighted Regression (GWR), is a technique of spatial analysis introduced in the in the 90's in the field of geography (Brunsdon et al., 1996). Unlike the classical regression model where the estimated coefficients are constant in space, GWR can model relationships that vary spatially between the independent variables and the dependent variable.

In recent years there has been methodological development of GWR models due to the increasing demand for the application of spatial data models in several research areas. Therefore, GWR has been adapted to problem-solve the basic model, namely, collinearity, the nature of the error term, spatial heteroskedasticity and some focus on different widths band to estimate the terms of regression.

In addition to these extensions, GWR has also expanded its application to a wide range of research areas, including statistics, economics, physics, engineering and health sciences (among others).

The aim of this study was to improve the understanding of the structure, contents, and trend of topics within the existing literature in the field of GWR. Conventional bibliometric methods and topic modeling techniques were performed. On the other hand latent Dirichlet allocation was used to identify patterns and trends.

2. METHOD

The methods and general procedure used during the research are depicted in Figure 1.

Initially, a systematic and reproducible approach is applied to gather and filter the data sample. Afterward, bibliometrics analysis is applied to investigate the collected data sample. Finally, a literature review is conducted to address and reveal research topics within the field of GWR models, for it we apply an unsupervised machine learning algorithm called latent Dirichlet allocation (LDA; Blei et al. 2003),

2.1 Dataset preparation

The search for articles was carried out through the Scopus, we decided to select this database since it is one of the databases most used by researchers (Harzing & Alakangas, 2016. Inclusion criteria focused on the selection of research articles containing information discussing the technicalities of GWR and/or its extensions, published in the English language in the period from October 1996 (date of first publication) to December 2019, also only peer-reviewed articles were considered. Searching was conducted on 29 June 2020 using the following query:

```
TITLE-ABS-KEY ("Geographically Weighted regression" OR "Heteroscedastic GWR" OR "Geographically Weighted Generalized Linear Modeling" OR "Geographically Weighted Logistic Regression" OR "Geographically Weighted Poisson Regression" OR "Geographically Weighted Multinomial") AND ( LIMIT-TO(DOCTYPE,"ar")) AND (LIMIT-TO( LANGUAGE, "English")) AND ( EXCLUDE (PUBYEAR, 2021)OR EXCLUDE(PUBYEAR,2020))
```

2.2 Bibliometrics and research mapping

Data processing in this part of the study was carried out using *bibliometrix* (Aria &

Cuccurullo, 2017), an opensource package to R programming language (R Core Team 2019), while for visualization of the co-authorship networks, the VOSviewer (Van Eck & Waltman, 2010) version 1.6.15 program was used.

2.3 Identifying research topics

To uncover latent topics, the topic model method latent Dirichlet allocation (LDA) (Blei, 2012; Blei et al., 2003) was used. LDA is based on Bayesian models and is considered an extension of Probabilistic Latent Semantic Analysis (Blei et al., 2003; Grün & Hornik, 2011). Figure 2 show the graphic model for LDA.

We decided to use the texts of the complete articles since, in contrast to the exclusive use of abstracts, it has been shown that this increases the quality of the topics and provides greater detail of the latent themes that can be found in a document (Syed & Spruit ,2017).

The procedure for the identification of topics through LDA, was divided into three stages: (i) preprocessing, (ii) creation of LDA model and (iii) labelling topics (Figure 3).

Preprocessing texts

Initially the collected documents were converted from pdf format to txt. Afterwards, numbers, punctuation marks and blank spaces were eliminated. A standard list of words called stop-words were also eliminated. In addition, all the words were converted to lowercase and, stemming was applied. Once the preprocessing phase was finished, the document terms matrix (*dtm*) was created. All stages of the preprocessing were made in the text mining package tm (Feinerer & Hornik, 2019), Software R (R Core Team, 2019).

Creation model Latent Dirichlet allocation

Table 1 shows the parameters used in the experimental configuration of the LDA model used in this article. Topic models are latent variable models of documents that use correlations between words and latent semantic themes in a collection of documents (Blei & Lafferty, 2007). This definition assumes that the expected number of topics *k* (i.e. latent variables) must be established *a priori*. Simulations were carried out varying k from 4 to 30 in incremental steps of two, an inference algorithm with 500 iterations was used, namely Gibbs sampling (Geman & Geman, 1984). Also, default value from the "textmineR" (Jones, 2019) package were used for Dirichlet parameters α , and ß was estimated based on the corpus. Quality LDA model was determined by utilizing a topic coherence measure (Röder et al., 2015).

Labelling topics

The topics are not semantically labeled for the LDA model and algorithmic analyzes are very limited in their ability to understand latent meanings of human language, so manual labeling is considered a standard in topic modeling (Lau et al., 2011). In addition, to improve the labeling of the topics, we visualized them in a two-dimensional area by computing the distance between topics (Chuang et al., 2012) by means of a multidimensional scaling analysis (Siever & Shirley, 2014).

2.4 Quantitative indices used to analyze the trend of topics

Due to the large number of articles and, therefore, the quantity of words, it is difficult to understand the topics and their trends intuitively. Therefore, we used some quantitative indices proposed by Xiong et al. (2019), which are obtained by adding documents-topic and topic-words distributions in order to make clear the results and findings. The description of the indexes is as follows. The distribution of topics over time is obtained by

$$\theta_k^y = \frac{\sum_{m\epsilon y} \theta_{mk}}{n^y} \quad (2)$$

where $m\epsilon y$ represents articles published in a given year, θ_{mk} the proportion of the k-th topic in each item and $n^y$the total number of articles published in the year (Xiong et al. 2019).

Topic distribution across journals is defined as the ratio of the k-th topic in the journal j:

θ_k^j

$$\theta_k^j = \frac{\sum_{m\epsilon j} \theta_{mk}}{n^j} \quad (3)$$

where, $m\epsilon j$ represents the articles in a particular journal, θ_{mk} the proportion of the k-th topic on each item, and n^j the total number of articles published in the journal j.

With the purpose of facilitating the characterization of the topics in terms of their tendency, we used simple regression slopes for each topic where the year was the dependent variable and the proportion of the topics in the corresponding year was the response variable (Griffiths and Steyvers 2004). The slopes obtained by regression were positive or negative at a statistical significance level of 0.05 and topics were classified as positive or negative trends respectively.

In order to quantify the activity and influence of the topics, we calculated the popularity of topics. The popularity of a topic, taking into account trends and the probability of the topic, can be described quantitatively as follows:

$$P^i = S^i_{NP} + S^i_{Tr} \quad (4)$$

$$S^i_{NP} = P^i_A / P^{max}_A \quad (5)$$

where P^i is the popularity of a topic i, S^i_{NP} is the normalized probability and S^i_{Tr} takes values of 1 if a topic shows a positive trend, 0.67 (fluctuating) and 0.33 (negative trend) (Xiong et al. 2019).

3. RESULTS AND DISCUSSION

3.1 Bibliometric analysis

The summary generated includes basic statistics about the analyzed dataset is presented in Table 2. Documents stemmed from 574 different journal and were published in the course of 24 years. A total of 3681 authors were involved in the scientific production on GWR. Among the papers, 103 were created by a single author whereas the overall Collaboration Index of the sample equals 2.51. This indicator denotes the average number of co-authors noted solely in multi-authored publications (Elango & Rajendran 2012).

In the following 15 years, after the first publication, the annual number of publications was less than 50, but since 2014 articles increased significantly and the annual publication was greater than 100 (Figure 4), also 70% of the articles have been published since that year. This grown could be due to introduction of new journals (Sun & Yin 2017). The compound annual growth rate was 28.31%.

Figure 4 (blue line) shows the distribution of average citations per year, two peaks stand out. The first refers to the first publication, while the second refers to three publications that appeared in 1998, which received an annual average of 42.25 and 16.81 cites, respectively. The falloff of citations does not imply that no important contributions appeared, but rather that it always needs about one decade for articles to be widely cited and recognized (Pilkington & Meredith 2009). The increase in the number of publications and its own acceleration of this growth in recent years seems to indicate that the use of GWR has not yet reached its full maturity.

Table 3 provides the list of top-twenty the most-cited authors. Of all the authors in the 1539 articles, the three most cited (based on local citations i.e number of citations included in the analyzed in this document) are Fotheringham AS, Charlton M, and Brunsdon C, All of them exceed 3,000 citations, while the remaining in all cases are less than 605.

In addition to focusing on the number of citations, we also inspected the most important papers, in terms of global citation (number of citations in the entire database Scopus) and local citations. The results are presented in Table 4 and sorted by local citations.

Table 5 shows, the most influential journals in terms of article count, these journals are distributed in different subject areas, such as environmental science, engineering, health, science, economics, social sciences and so on. This shows that the theory and method of GWR is widely used, and it has attracted a lot of attention from scholars in different disciplines. the journals with the largest published articles are *Applied Geography*, *International Journal of Environmental Research and Public Health*, and *Sustainability* (Switzerland). while the most cited were *Geographical Analysis* with 711 citations, followed by *Environment And Planning A* (657 citations).

In terms of publication production by country (in these cases, if an article had two authors from different countries, the article was counted twice), articles analyzed originated from institutions linked to 87 countries. The map (Figure 5) show the geographic distribution of GWR publications, it can be seen that are dominated by authors located in a few geographical regions. China and USA leads with this regard with 1,181 and 1,045 articles respectively. United Kingdom (UK), Canada and Australia appear in the third, fourth, and fifth position, with 200, 140, 117 articles respectively.

As in most other research fields, United States and China were the largest contributor, China also showing remarkable overall growth in research output during the last decade (Aksnes et al., 2014); this growth is also evident in articles about GWR.

Figure 6 shows, 940 links (of the [87* 86 / 2] = 3741 possible links) between nations. Thus, it was possible to observe that international collaboration is heavily concentrated in USA, China, and UK. However China is the main partner of USA in terms of international co-authorship on GWR-related publications.

Visual inspection of the map shows that from the Africa countries only South Africa, Egypt, Zimbabwe ,and Ghana contributes, while Latin America participation is limited mainly to Brazil, Argentina, Chile, Colombia, Ecuador ,and Mexico. It should be noted that Latin American countries such as Venezuela, Ecuador, Cuba, Trinidad y Tobago among others were excluded due to the restriction in published articles (minimum 3).

In a network map, two agents are positioned close to each other if they communicate intensively, but not on the basis of fixed geographical coordinates, likewise two nations may not relate intensively, but may share a common pattern of relations with third parties (Leydesdorff et al., 2013). Figure 7 shows the network of international coauthorship relations among 57 countries, since to improve the visualization of the network it was decided to include countries with 3 or more published articles, also in this visualization was used fractional counting. The circle node represents the country, and the connection represents the partnership. The thickness of the line reflects the intensity of cooperation. The thicker the connection between the two nodes, the closer and more frequent the cooperation between the two countries. There are nine colors in Figure 7, indicating that these 57 countries are grouped into nine clusters. There is a stronger cooperative relationship between countries in the same cluster than between countries in different clusters. Of course, this does not mean that there is no cooperation between countries in different clusters rather, there may be some common research topics among the countries of the same cluster, making their cooperation closer. We observed that countries such as the Philippines, Bangladesh, Malaysia and Indonesia have international co-authorship relationships on GWR related issues with industrialized neighbors in the Asia-Pacific region, the US and some European nations. Many of these relationships can be a consequence of scholars having studied abroad as masters and Ph.D. students.

This reflects the pattern in science more generally (Wagner et al., 2015), China has emerged as a prominent American contributor, surpassing the historical ties of collaboration in other areas between European and North American countries.

Given the origin of the authors, special attention is paid to Latin America and the European Union (EU) (specially Spain and Portugal). Figure 8 provides the collaboration network among these nations and this shows the much stronger connection between Germany, Netherland and Austria. On the other hand we observe that Spain acts as a connection node between the EU and Latin America more than Portugal, also Brazil is the country with the greatest integration in terms of co-authorship with EU countries.

3.2 Topic modeling analysis

Table 6 provides an overview of the complete set of articles or corpus used in this study. The top 25 terms (unigram and bigram) with high occurrence frequency are listed in Table 7. We observed that the words found in the articles considered represent the variety of topics investigated in the field of GWR.

Table 8 presents the 20 topics estimated by our model and for each of them the 15 most common terms and topic label organized from most prevalent to least prevalent. Figure 9 provides an additional representation of the 20 uncovered GWR topics, alongside their proportions across the entire corpus. Here, the distance between the nodes represents the topic similarity with respect to the distribution of words, whilst the size of the nodes indicates the topic prevalence within the corpus, with larger nodes representing topics being more prominent within the corpus. Thus, we can see, for instance, that the topic “Spatial variation”, “Spatial autocorrelation”, GWR model”, and “Model estimation” were the most prominent of the 20 topics identified across the entire corpus and so on. In addition, these four topics are grouped into a cluster, which could be indicating the use of common specific words within this topic.

Trends of topics

We identified the tendency of each of the 20 topics over time. We found that the probabilities of 7 of them “China air quality”, “Education”, “Health”, “House price”, “Land use”, “Precipitation”, and “Remote sense” increased progressively over time, in three of them decreased “GWR model”, “Model estimation”, and “Spatial variation” (Figure 10) while the remainder fluctuated over time, without prominent trends. The results found are consistent with the idea that research shows strong trends with topics that rise and fall regularly in prevalence (Griffiths & Steyvers, 2004).

The growing trend of the topic "China air quality" is due to the fact that in recent decades, China's rapid development of economics has caused all kinds of pollution problems. Air pollution is the most serious environmental problem in China, which has imposed about 5% Country's GDP to suspend gaseous particulates and caused about 3 million Chinese deaths in 2012 (Langrish et al., 2012).

The increase in the proportion of some topics, indicates that these are emerging fields of research, while its decrease shows a trend of lower research interest. In addition, in some topic the high frequency pattern found was followed by a negative trend during the period of study, indicating a possible decrease in their popularity within the scientific community.

The obtained results for the distribution of the topics by year are also is presented by a heatmap (Figure 11). Thus, the color of the pixel represents the probability that a certain topic is mentioned in a particular year. So, and especially in the last five years , have a relatively broad scope, while in the early years focus on more specific topics especially those topics related to the development of the model.

Taking into account the journal, the results for 25 of them are grouped according to the topics addressed by them (Figure 12), there is some similar prevalence distribution of topics for example *Accident Analysis and Prevention* (Accid. Anal. Prev.) and *Journal of Transport Geography (*J. Transp. Geogr.), which were focused on topics “Transport” and “Spatial variation”.

Popularity topics

The popularity of the topics, taking into account the trend and the probability of the topic, are presented in Table 9. It is observed that the five most popular topics in descending order are; “GWR models”, “Health care”, “Precipitation”**, “**Spatial variation”, and “Land use”. It can be seen that although "Health" and "Precipitation" have average probabilities, they occupy the first places in popularity. In general, topics with increasing trends with high and moderate average probabilities have higher popularity scores, while some topic with moderate probabilities and increasing trends improved considerably.

The topics with the lowest popularity scores were “Soil”, “Transport”, “Water quality” they did not showed a marked trend and had the low average probabilities. Previous results show that the GWR technique has been used by researchers who are focused on current concerns such as environmental pollution, health and socio-economic issues such as poverty. According to Cao et al., (2009) and Zhong et al., (2013) there are direct and indirect links between the listed topics; also, Beggs and Bambrick (2005) state that there are direct links between climate change and disease, and Cleaver and Schreiber (1994) state that growth population, agricultural performance and environmental degradation are also interrelated.

4. CONCLUSIONS

The research was focused on the period between October 1996 and to December 2019 using 1539 published articles in peer-reviewed journals. It should be noted that this study is limited by the exclusion of books, reviews, book chapters, gray literature and reports from the research, also, the data was collected only from the Scopus database and furthermore, only articles were taken into account.

China, USA, UK, Canada and Australia are the most productive countries/territories. The international collaboration is heavily concentrated in USA, China, and UK. China is the main partner of USA in terms of international co-authorship on GWR-related publications.

There were not many articles published by developing countries in the field of GWR. Latin America participation is limited, Spain acts as a connection node between the EU and Latin America more than Portugal. Brazil is the country with the greatest integration in terms of co-authorship with EU countries.

With this paper, we also contribute to the methodology related to literature surveys and bibliometric analyses. Notably, we utilize a more advanced methodology beyond traditional bibliometric analyses. We conducted topic modelling to extract the latent topics from the collected literature. We quantitatively analyzed and identified 20 latent topics and their trends. The data was presented in the form of topic distributions, which revealed how the scope of research in GWR varies over time. However, as a limitation it was found that LDA cannot automatically assign each topic to a specific subfield. The distribution of topics was wide and relatively uniform over the last seven years, while the first years, as expected, presented the opposite. Thus, this means that the interests of research and applications of the technique have varied significantly.

REFERENCES

Aksnes, D. W., van Leeuwen, T. N., & Sivertsen, G. (2014). The effect of booming countries on changes in the relative specialization index (RSI) on country level. *Scientometrics*, *101*(2), 1391-1401. https://doi.org/10.1007/s11192-014-1245-3

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics*, *11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2012). On Smoothing and Inference for Topic Models. *In UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 27–34). Montreal, Quebec, Canada: AUAI Press Arlington.

Beggs, P. J., & Bambrick, H. J. (2005). Is the global rise of asthma an early impact of anthropogenic climate change?. *Environmental health perspectives*, *113*(8), 915-919. https://doi.org/10.1289/ehp.7724

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, *1*(1), 121-143. http://doi.org/10.1214/06-BA104

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35. http://doi.org/10.1214/07-AOAS114

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, *28*(4), 281-298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Cao, S., Zhong, B., Yue, H., Zeng, H., & Zeng, J. (2009). Development and testing of a sustainable environmental restoration policy on eradicating the poverty trap in China's Changting County. *Proceedings of the National Academy of Sciences*, *106*(26), 10712-10716. https://doi.org/10.1073/pnas.0900197106

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012, May). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443-452). https://doi.org/10.1145/2207676.2207738

Elango, B., & Rajendran, P. (2012). Authorship trends and collaboration pattern in the marine sciences literature: a scientometric study. *International Journal of Information Dissemination and Technology*, *2*(3), 166-169.

Feinerer. I., and K. Hornik. (2019). "Text Mining Package." Available at: https://cran.r-project.org/web/packages/tm (Accessed: 22 Jun 2020).

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741. https://doi.org/10.1109/TPAMI.1984.4767596

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787-804. https://doi.org/10.1007/s11192-015-1798-9

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software*, *40*(13), 1-30. https://doi.org/10.18637/jss.v040.i13

Jones, T. (2019). textmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.4. https://CRAN.R-project.org/package=textmineR

Langrish, J. P., Bosson, J., Unosson, J., Muala, A., Newby, D. E., Mills, N. L., ... & Sandström, T. (2012). Cardiovascular effects of particulate air pollution exposure: time course and underlying mechanisms. *Journal of internal medicine*, *272*(3), 224-239. https://doi.org/10.1111/j.1365-2796.2012.02566.x

Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011, June). Automatic labelling of topic models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 1536-1545).

Leydesdorff, L., Wagner, C., Park, H. W., & Adams, J. (2013). International collaboration in science: The global map and the network. *arXiv preprint arXiv:1301.0801*.

Pilkington, A., & Meredith, J. (2009). The evolution of the intellectual structure of operations management—1980–2006: A citation/co-citation analysis. *Journal of operations management*, *27*(3), 185-202. https://doi.org/10.1016/j.jom.2008.08.001

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008, August). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 569-577). https://doi.org/10.1145/1401890.1401960

Team, R. C. (2019). R: A language and environment for statistical computing. Available at: https://www.r-project.org.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408). https://doi.org/10.1145/2684822.2685324

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*, *77*, 49-66. https://doi.org/10.1016/j.trc.2017.01.013

Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science* *and advanced analytics (DSAA)* (pp. 165-174). IEEE. https://doi.org/10.1109/DSAA.2017.61

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *scientometrics*, *84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

Wagner, C. S., Bornmann, L., & Leydesdorff, L. (2015). Recent developments in China–US cooperation in science. *Minerva*, *53*(3), 199-214. https://doi.org/10.1007/s11024-015-9273-6

Wang, C., Paisley, J., & Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 752-760). JMLR Workshop and Conference Proceedings.

Wang, W., Xu, S., & Yan, T. (2018). Structure identification and model selection in geographically weighted quantile regression models. *Spatial Statistics*, *26*, 21-37. https://doi.org/10.1016/j.spasta.2018.05.003

Xiong, H., Cheng, Y., Zhao, W., & Liu, J. (2019). Analyzing scientific research topics in manufacturing field using a topic model. *Computers & Industrial Engineering*, *135*, 333-347. https://doi.org/10.1016/j.cie.2019.06.010

Zhong, B., Peng, S., Zhang, Q., Ma, H., & Cao, S. (2013). Using an ecological economics approach to support the restoration of collapsing gullies in southern China. *Land Use Policy*, *32*, 119-124. https://doi.org/10.1016/j.landusepol.2012.10.005

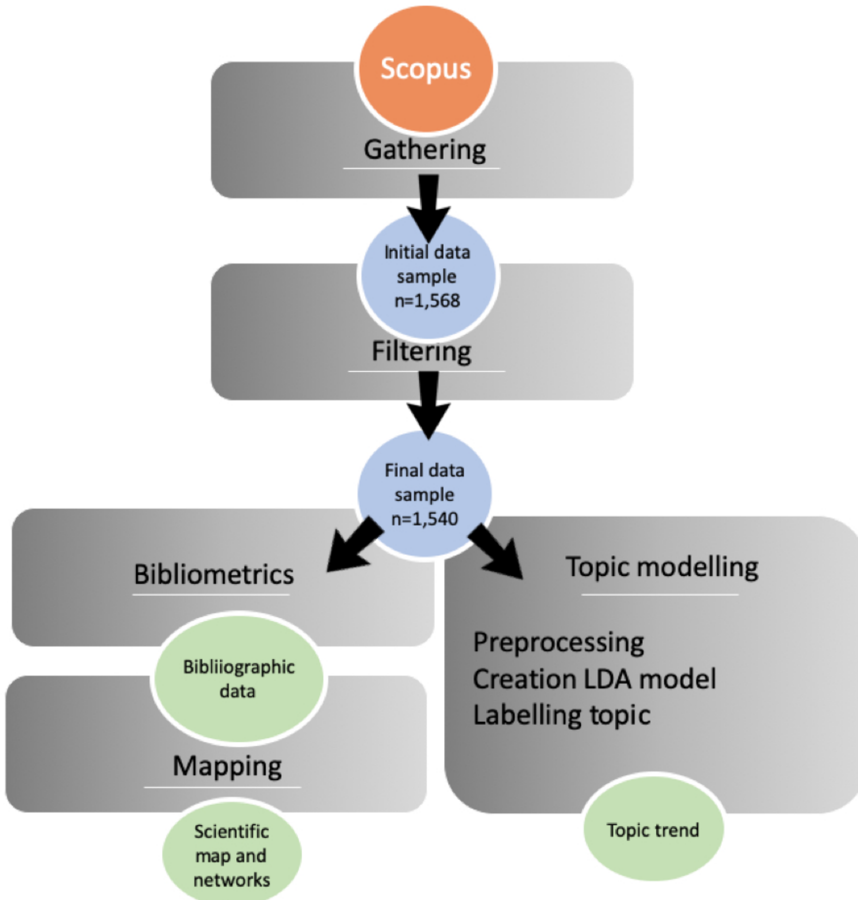

Figure 1. Methods and general procedure used in the performed study

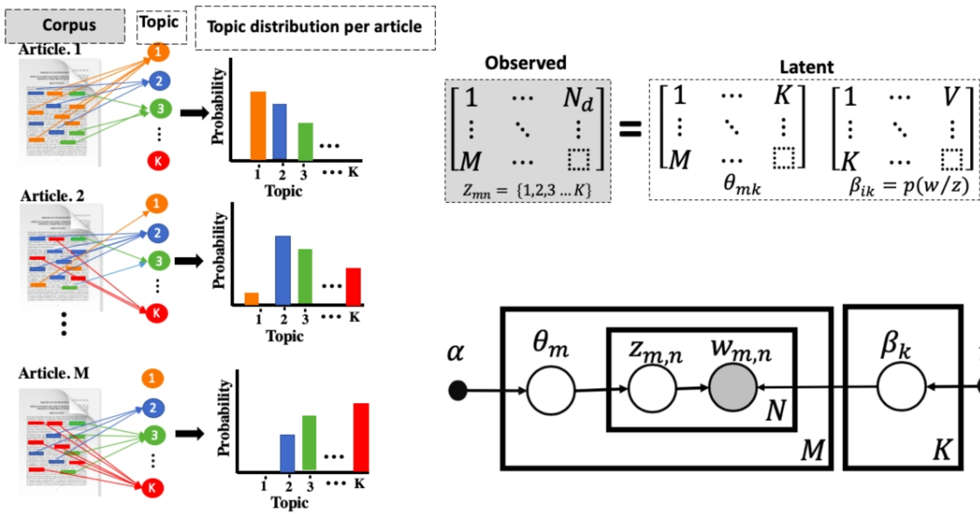

Fig. 2 Graphic model for Latent Dirichlet Allocation LDA (Source: adapted from Blei et al. 2003). K, M, N and V denote the number of topics, number of article, words in article and words in the vocabulary respectively; �� and �� (Dirichlet hyper-parameters) are parameters of the prior distributions over �� and �� respectively ; ��m is the distribution of topics for article m (real vector of length ��) ; ��m,�� is the topic for the ����h word in the m-��h article; ��m,�� is the �� th word of the m-��h document; ���� is the distribution of words for topic �� (real vector of length ��).

428x220mm (72 x 72 DPI)

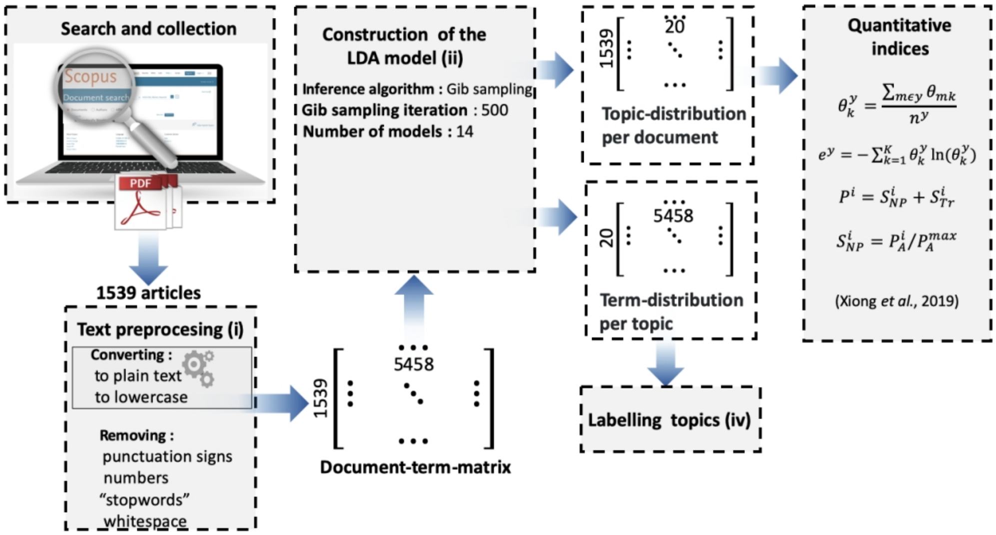

Figure 3. Scheme of the methodological process used in the identification of research topics in Geographically Weighted Regression through latent Dirichlet allocation

493x267mm (72 x 72 DPI)

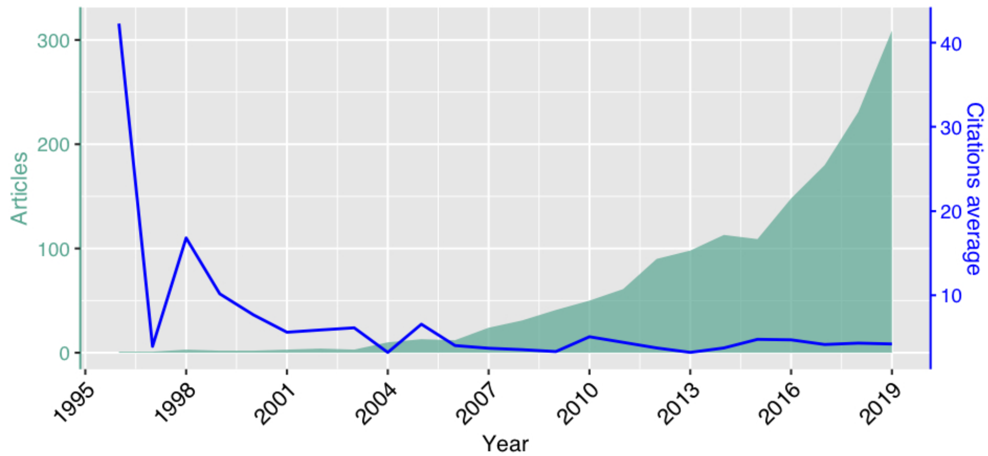

Figure 4. Number of articles an citation average (blue line) in the years 1996–2019 based on the analyzed dataset.

336x158mm (72 x 72 DPI)

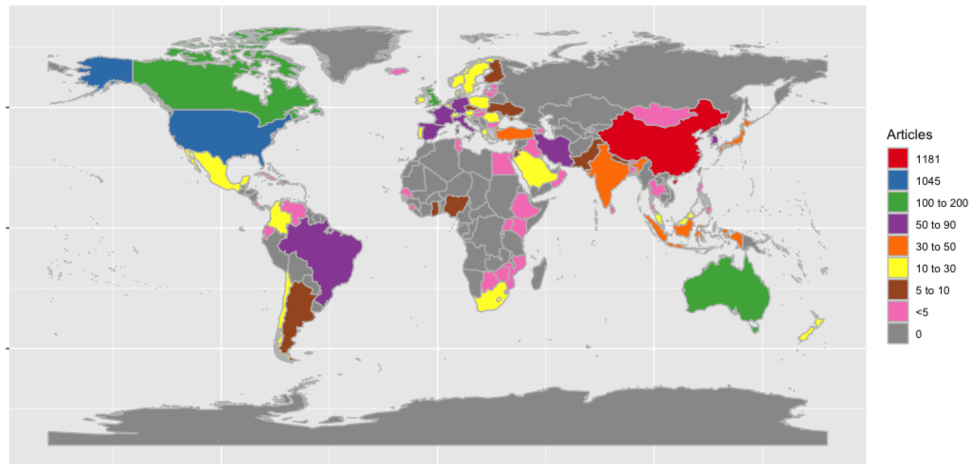

Figure 5. Global geographic distribution of GWR publications in the years 1996–2019

371x183mm (72 x 72 DPI)

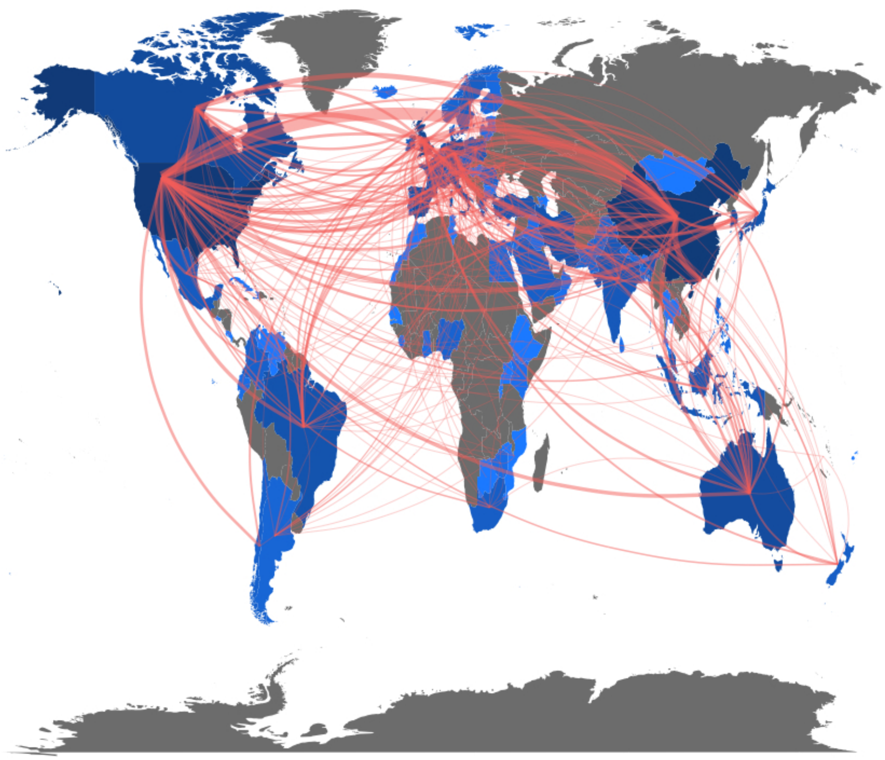

Figure 6. Map of international collaborations

327x291mm (72 x 72 DPI)

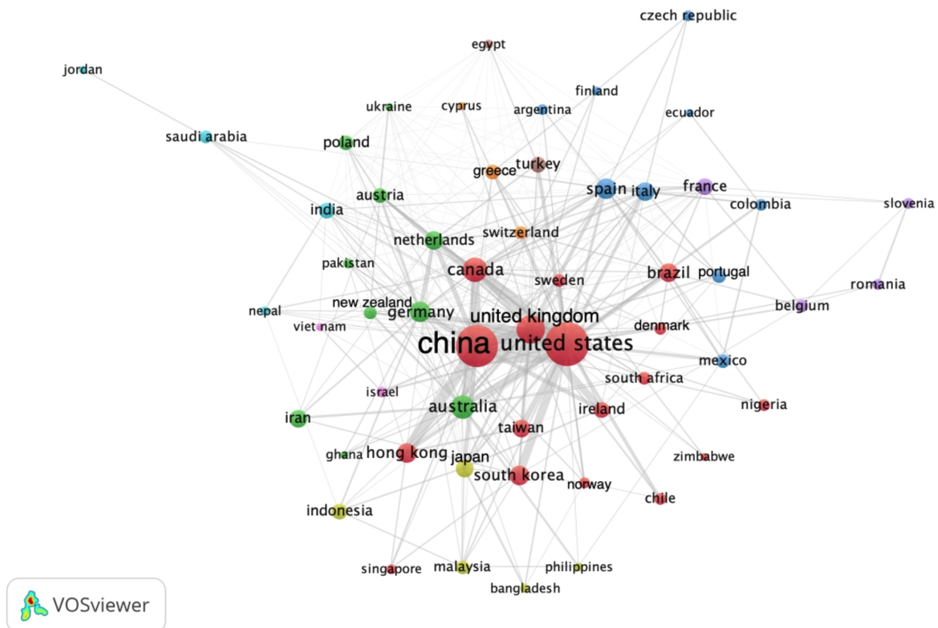

Figure 7. International co-authorship relationships of 57 countries with more than 3 documents on GWR between 1996 and 2019

721x496mm (72 x 72 DPI)

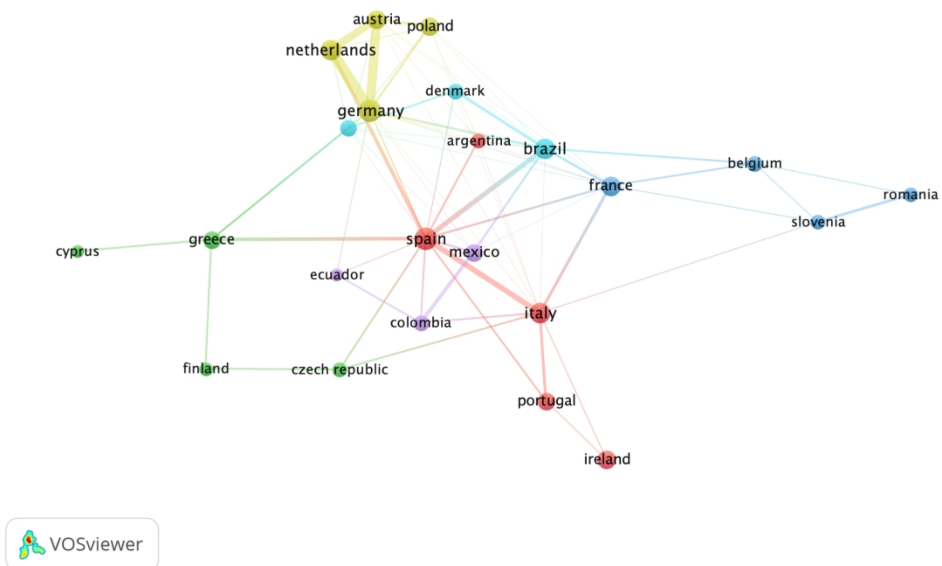

Figure 8. Co-authorship network between countries of the European Union and Latin America

721x496mm (72 x 72 DPI)

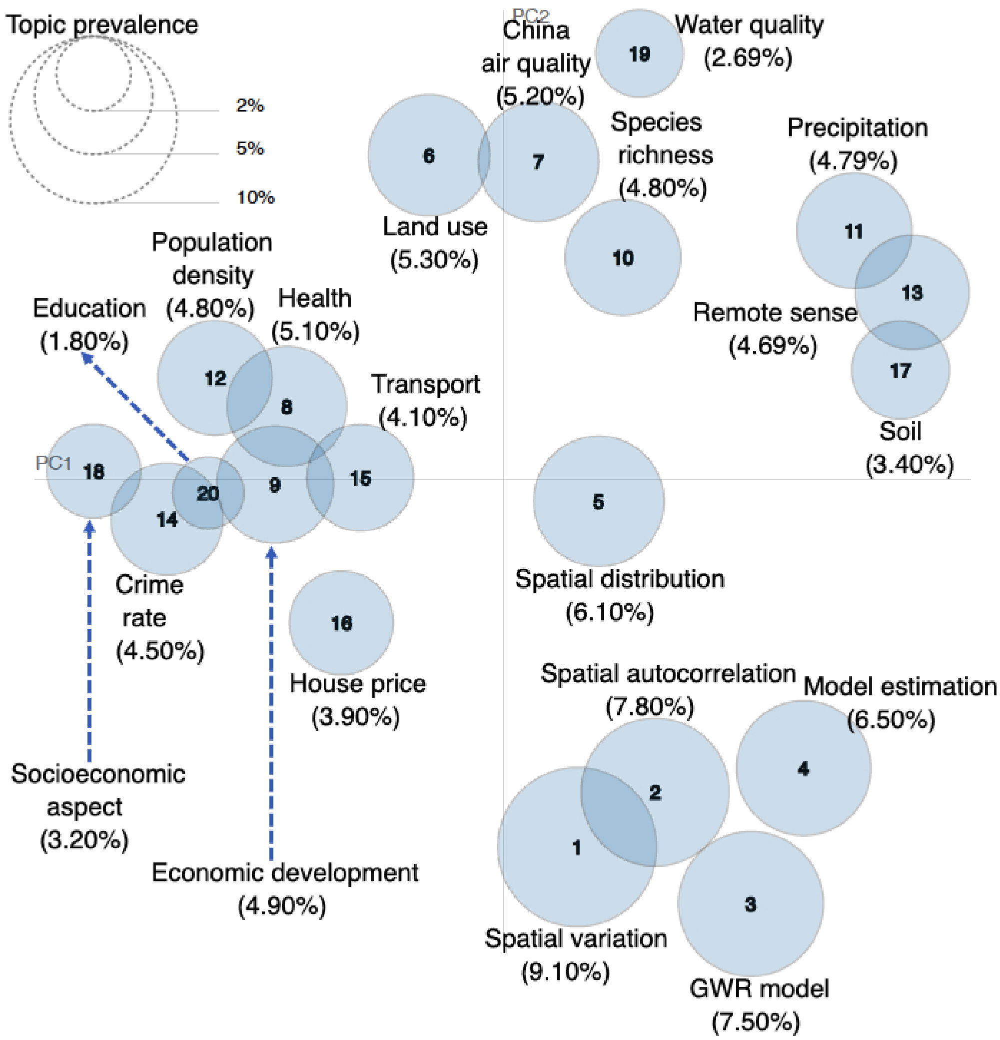

Figure 9. Inter-topic distance map that shows a two-dimensional representation via multidimensional scaling (all nodes add up to 100%).

246x277mm (72 x 72 DPI)

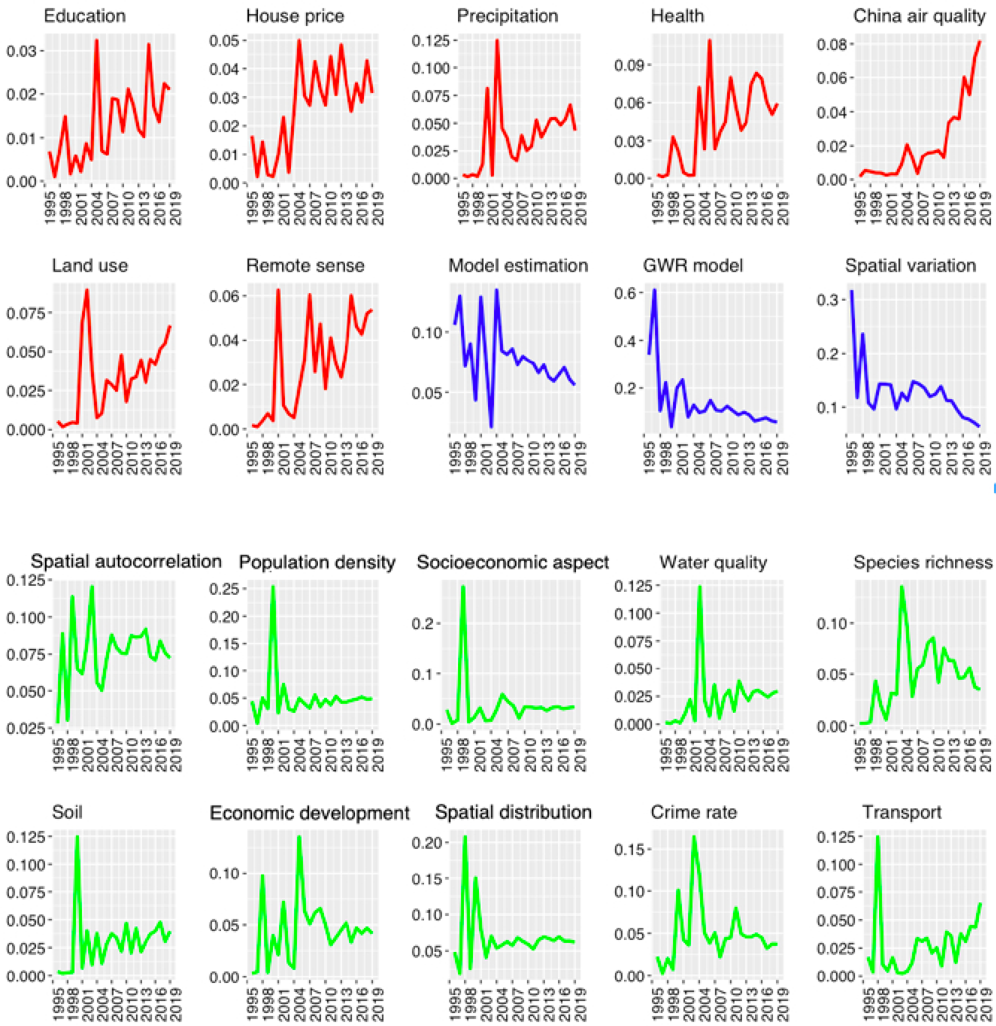

Figure 10. Topic trends research in Geographically Weighted Regression during Octuber 1996– December 2019. The red color indicates topics with increasing tendency, blue with decreasing tendency and green fluctuating

236x243mm (72 x 72 DPI)

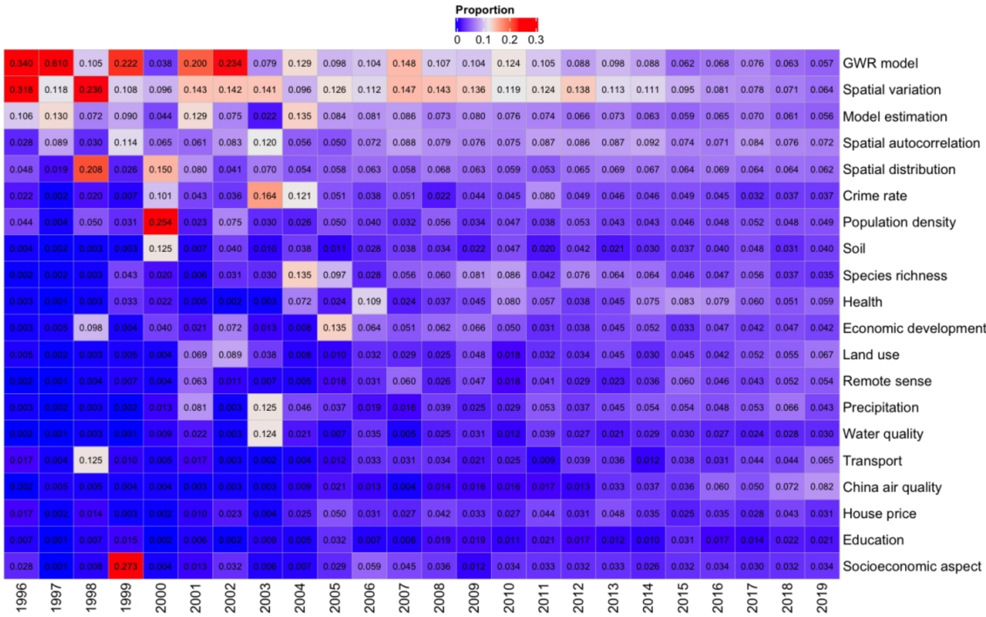

Figure 11. Heatmap overview of the proportional topic in the 24 year analysed. Values are in percentages, and column totals sum up to 100%.

508x317mm (72 x 72 DPI)

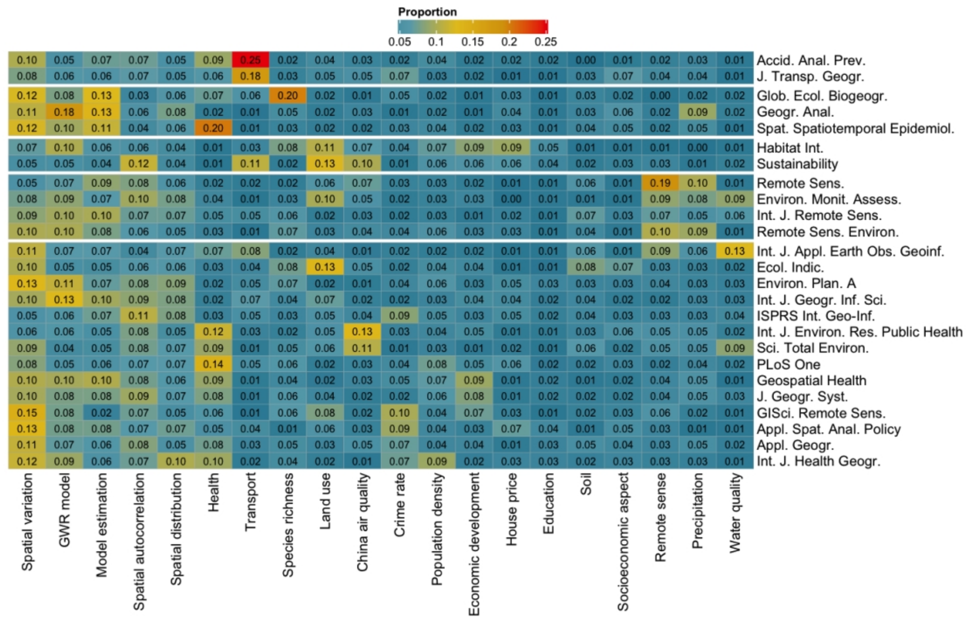

Figure 12. Heatmap overview of the proportional topic in the 25 analyzed journals. Values are in percentages, and row totals sum up to 100%

508x317mm (72 x 72 DPI)

Table 1. Experimental settings Latent Dirichlet Allocation (LDA) parameters

Component	Candidate
Inference algorithm	Gibbs sampling
The number of topics K	4:30 by 2
Gibbs sampling iteration	500
Dirichlet parameter α	0.1 (default value of the FitLdaModel function in the *texmineR* package)
Dirichlet parameter β	Estimated from the corpus

Table 2. Summary of the bibliometric data sample.

Description	Results
Timespan	1996:2019 (24 years)
Sources	574
Documents	1539
Average years from publication	4.95
Average citations per documents	22.06
Average citations per year per doc	3.217
References	70393
Authors	3681
Author Appearances	5819
Authors of single-authored documents	103
Authors of multi-authored documents	3578
Single-authored documents	113
Documents per Author	0.418
Authors per Document	2.39
Co-Authors per Documents	3.78
Collaboration Index	2.51

Table 3. Top 20 cited authors.

Authors	Articles	Citation	Articles Fractionalized
FOTHERINGHAM AS	32	3435	8.33
CHARLTON M	20	3152	4.38
BRUNSDON C	25	3046	7.30
LIU Y	36	604	7.86
WANG Y	30	488	6.74
ZHANG L	31	449	8.53
ZHANG Y	23	374	4.53
LI X	19	333	3.92
HARRIS P	16	322	5.41
WANG J	18	311	3.86
LI Y	16	266	3.34
LIU J	16	247	3.36
LI W	26	224	6.61
LI Z	17	210	3.41
ZHANG C	18	200	4.18
WANG L	19	197	4.59
ZHANG H	17	193	3.80
LI H	16	184	4.44
WANG S	16	158	3.26
ZHAO Y	19	106	3.39

Table 4. top ten most cited articles

Title	Author(s)	Year	Source	Local citations	Global citations
Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity.	Brunsdon, C., A. S. Fotheringham, and M. E. Charlton.	1996	Geographical Analysis	416	1014
Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis.	Fotheringham, A. S., Charlton, M. E., & Brunsdon, C.	1998	Environment and Planning A,	200	469
Multicollinearity and correlation among local regression coefficients in geographically weighted regression	Wheeler, D., and Tiefelsdorf,	2005	Journal of Geographical Systems	185	381
Geographically weighted Poisson regression for disease association mapping	Nakaya, T., Fotheringham, A. S., Brunsdon, C., and Charlton, M.	2005	Statistics in Medicine	116	244
Statistical Tests for Spatial Nonstationarity Based on the Geographically Weighted Regression Model	Leung, Y., Mei, C.-L., and Zhang, W.-X	2000	Environment and Planning A	104	227
Some Notes on Parametric Significance Tests for Geographically Weighted Regression	Brunsdon, C., Fotheringham, A. S., and Charlton, M.	1999	Journal of Regional Science	96	230
Geographical weighting as a further refinement to regression modelling: An example focused on the NDVI–rainfall relationship	Foody, G.	2003	Remote Sensing of Environment	85	202
Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation	TU, J., and XIA, Z.	2008	Science of The Total Environment	74	196
Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression	Wheeler, D. C.	2007	Environment and Planning A	73	124
Mapping the Results of Geographically Weighted Regression	Mennis, J.	2006	The Cartographic Journal	67	124

Table 5. Most important journals in terms of article count.

Source	**Abbreviation**	Articles	Cites	Year start
Applied Geography	Appl. Geogr.	58	484	2009
International Journal of Environmental Research and Public Health	Int. J. Environ. Res. Public Health	41	46	2011
Sustainability (Switzerland)	Sustainability	39	195	2014
Remote Sensing	Remote Sens.	34	71	2014
Isprs International Journal of Geo-Information	ISPRS Int. Geo-Inf.	32	20	2014
Plos One	PLoS One	29	261	2011
Science of the Total Environment	Sci. Total Environ.	27	130	2008
Remote Sensing of Environment	Remote Sens. Environ.	21	466	2003
International Journal Of Health Geographics	Int. J. Health Geogr.	20	95	2009
Journal of Transport Geography	J. Transp. Geogr.	19	172	2006
Environment and Planning A	Environ. Plan. A	18	657	1998
Giscience and Remote Sensing	GISci. Remote Sens.	17	33	2004
Applied Spatial Analysis and Policy	Appl. Spat. Anal. Policy	16	48	2009
Ecological Indicators	Ecol. Indic.	16	61	2011
Geographical Analysis	Geogr. Anal.	16	711	1996
International Journal of Geographical Information Science	Int. J. Geogr. Inf. Sci.	16	209	2010
Accident Analysis and Prevention	Accid. Anal. Prev.	14	90	2010
Spatial and Spatio-Temporal Epidemiology	Spat. Spatiotemporal Epidemiol.	14	13	2009
Journal of Geographical Systems	J. Geogr. Syst.	13	338	2002
International Journal of Remote Sensing	Int. J. Remote Sens.	12	323	2005
Geospatial Health	Geospatial Health	11	28	2010
International Journal of Applied Earth Observation And Geoinformation	Int. J. Appl. Earth Obs. Geoinf.	11	63	2011
Environmental Monitoring and Assessment	Environ. Monit. Assess.	10	33	2014
Global Ecology and Biogeography	Glob. Ecol. Biogeogr.	10	224	2004
Habitat International	Habitat Int.	10	85	2014

Table 6. Overview of the corpus : year in which the article was published; N, the number of document; N(%), percentaje of the number of articles in relation to the total articles; $\overline{W}$, mean number of words within each article; Std.W, estimated standard deviation of words within each article; $\overline{V}$, mean vocabulary (number of unique words) within each article; Std.V, estimated standard deviation of unique words within each article

Year	N	N (%)	$\overline{W}$	Std W	$\overline{V}$	Std V
1996	1	0,06	8669,0	-	1757,0	-
1997	1	0,06	15020,0	-	1825,0	-
1998	3	0,19	10046,0	2047,8	2027,5	593,3
1999	2	0,13	9816,0	3740,6	1987,5	758,7
2000	2	0,13	8418,0	5925,6	1617,5	815,3
2001	3	0,19	9456,3	1219,9	2017,3	142,4
2002	4	0,26	10338,5	2401,6	1929,8	257,0
2003	3	0,19	11715,3	1739,9	2341,7	208,4
2004	10	0,65	9564,0	1096,1	2017,5	259,7
2005	13	0,84	10856,5	3528,2	2194,4	582,4
2006	12	0,78	9258,4	1926,7	1921,3	342,9
2007	24	1,56	9562,3	2707,6	2001,7	436,4
2008	31	2,01	9978,1	2303,5	2122,7	414,8
2009	41	2,66	10270,7	3646,2	2010,0	599,5
2010	50	3,25	10014,3	2526,5	2080,6	439,0
2011	61	3,96	10291,8	3113,5	2119,4	503,5
2012	90	5,85	9907,9	3418,3	2018,7	559,2
2013	98	6,37	9807,7	3438,2	2014,3	590,6
2014	113	7,34	10460,3	3158,0	2140,1	449,0
2015	109	7,08	10731,1	3370,8	2154,0	501,5
2016	148	9,62	11244,7	3572,3	2259,2	573,5
2017	180	11,70	11526,8	3358,8	2246,7	455,6
2018	231	15,01	12057,6	3538,3	2331,2	513,4
2019	309	20,08	11393,0	3594,1	2216,6	523,0
Total	1539					

Table 7. Top 25 most frequent terms (unigram and bigram in the complete corpus. idf is the inverse document frequency

Rank	Term	Term frequency	Document frequency	idf	Rank	Term	Term frequency	Document frequency	idf
1	spatial	77537	1531	0.0052	28	geographically_weight	14522	1527	0.0078
2	variable	48528	1524	0.0098	32	weight_regression	13823	1522	0.0111
3	local	31216	1505	0.0223	35	gwr_model	13546	1267	0.1945
4	land	27264	1172	0.2724	74	regression_model	8561	1362	0.1222
5	estimate	26401	1495	0.0290	88	land_use	7731	776	0.6847
6	urban	26017	1212	0.2389	145	spatial_autocorrelation	5602	930	0.5037
7	relationship	25740	1511	0.0184	184	spatial_distribution	4923	999	0.4321
8	factor	23420	1442	0.0651	194	explanatory_variable	4731	854	0.5890
9	population	20669	1175	0.2699	199	independent_variable	4681	1016	0.4153
10	region	19289	1406	0.0904	202	spatial_variation	4663	1123	0.3151
11	level	19120	1460	0.0527	203	remote_sense	4656	575	0.9845
12	parameter	18307	1396	0.0975	231	dependent_variable	4055	1113	0.3241
13	coefficient	18273	1065	0.3682	239	spatial_pattern	3922	951	0.4814
14	table	18185	1480	0.0391	255	house_price	3750	288	1.6759
15	global	18140	1446	0.0623	258	population_density	3717	554	1.0217
16	china	16886	824	0.6247	259	ols_model	3712	640	0.8774
17	city	16818	1048	0.3842	260	parameter_estimate	3703	745	0.7255
18	distribution	16725	1468	0.0472	261	spatial_heterogeneity	3692	818	0.6320
19	change	16619	1419	0.0812	276	regression_gwr	3519	1406	0.0904
20	map	16218	1443	0.0644	282	land_cover	3506	452	1.2252
21	location	15903	1442	0.0651	295	spatially_vary	3273	1131	0.3080
22	numb	15752	1485	0.0357	318	spatial_nonstationarity	3033	933	0.5005
23	base	15594	1513	0.0170	329	remote_sens	2954	333	1.5307
24	increase	15251	1483	0.0371	331	spatial_analysis	2944	959	0.4730

25	pattern	15052	1412	0.0861	341	public_health	2858	461	1.2055

Table 8. 20 uncovered topics from 1539 Geographically Weighted Regression articles published in the period 1996–2019. Each topic displays the Top 15 high-frequency terms. The topics are manually labelled with a logical topic description that best captures the semantics of the top words.

Topic	**Prevalence**	**top_terms**
Spatial variation	9.10	local, relationship, variable, global, spatial, variation, parameter, estimate, vary, fotheringham, spatial_variation, geographically_weight, spatially, parameter_estimate, approach
Spatial autocorrelation	7.80	spatial, variable, ols, coefficient, gwr_model, autocorrelation, residual, local, spatial_autocorrelation, relationship, moran, factor, ols_model, distribution, table
GWR model	7.50	coefficient, bandwidth, estimate, spatial, test, parameter, function, kernel, geographically_weight, matrix, weight_regression, location, statistic, local, vary
Model estimation	6.50	spatial, error, variable, prediction, estimate, sample, approach, parameter, performance, linear, function, plot, numb, estimation, time
Spatial distribution	6.10	map, location, factor, distance, base, numb, cluster, process, apply, provide, logistic, main, probability, site, represent
Land use	5.30	urban, land, city, landscape, change, land_use, development, plan, pattern, urbanization, space, expansion, density, green, spatial
China air quality	5.20	china, emission, air, factor, pollution, city, energy, concentration, wang, province, environ, chinese, region, level, beijing
Health	5.10	health, risk, disease, cancer, mortality, rate, factor, incidence, care, association, cluster, population, prevalence, public_health, county
Economic development	4.90	regional, economic, region, growth, spatial, development, policy, county, province, industry, level, market, sector, economy, capital
Species richness	4.80	forest, species, tree, climate, richness, environmental, change, pattern, scale, relationship, ecology, diversity, ecosystem, global, ecological
Precipitation	4.79	precipitation, temperature, rainfall, climate, annual, average, station, elevation, region, season, monthly, air, meteorological, change, winter
Population density	4.70	population, rate, country, variable, change, municipality, level, social, age, increase, factor, migration, south, national, region
Remote sense	4.69	remote, sense, ndvi, vegetation, remote_sense, image, cover, surface, resolution, sens, land, remote_sens, satellite, map, pixel
Crime rate	4.50	crime, population, census, county, neighborhood, density, rate, city, spatial, social, tract, percent, american, level, block
Transport	4.10	road, transportation, travel, transport, variable, crash, numb, station, network, time, transit, distance, bus, trip, vehicle
House price	3.90	house, price, property, house_price, spatial, market, hedonic, residential, variable, real, estate, distance, city, urban, real_estate
Soil	3.40	soil, spatial, sample, map, organic, prediction, kriging, environmental, carbon, crop, variable, mlr, content, property, predict
Socioeconomic aspect	3.20	household, poverty, food, health, income, environment, physical, obesity, social, public, service, activity, age, individual, variable
Water quality	2.69	water, quality, river, land, lake, watershed, water_quality, surface, source, groundwater, site, basin, sample, metal, concentration
Education	1.80	school, district, education, variable, community, educational, facility, class, service, bank, factor, indicator, performance, level, birth

Table 9. Topic popularity

Rank	Topic	Probability	Normalized probability	Trend	Popularity
1	GWR model	0.1395	1.0000	0.33	1.3300
2	Health	0.0420	0.3012	1.00	1.3012
3	Precipitation	0.0374	0.2685	1.00	1.2685
4	Spatial variation	0.1274	0.9133	0.33	1.2433
5	Land use	0.0327	0.2347	1.00	1.2347
6	Spatial autocorrelation	0.0757	0.5425	0.67	1.2125
7	Remote sense	0.0286	0.2054	1.00	1.2054
8	House price	0.0263	0.1884	1.00	1.1884
9	Spatial distribution	0.0684	0.4904	0.67	1.1604
10	China air quality	0.0217	0.1554	1.00	1.1554
11	Education	0.0131	0.0939	1.00	1.0939
12	Population density	0.0506	0.3626	0.67	1.0326
13	Crime rate	0.0495	0.3549	0.67	1.0249
14	Species richness	0.0478	0.3427	0.67	1.0127
15	Economic development	0.0445	0.3193	0.67	0.9893
16	Socioeconomic aspect	0.0364	0.2613	0.67	0.9313
17	Model estimation	0.0778	0.5579	0.33	0.8879
18	Soil	0.0299	0.2146	0.67	0.8846
19	Transport	0.0275	0.1974	0.67	0.8674
20	Water quality	0.0232	0.1662	0.67	0.8362