IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2022, accepted February 2, 2022, date of publication February 10, 2022, date of current version February 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150869

Building a Gold Standard Dataset to Identify Articles About Geographic Information Science

CARLOS LÓPEZ-VÁZQUEZ[1], (Senior Member, IEEE), MARÍA ESTER GONZALEZ-CAMPOS[2], MIGUEL A. BERNABÉ-POVEDA[3], DANIELA MOCTEZUMA[4], ESTHER HOCHSZTAIN[5], MARÍA A. BARRERA[6], CARLOS GRANELL-CANUT[7], MARÍA F. LEÓN-PAZMIÑO[8], PABLO LÓPEZ-RAMÍREZ[4], VILLIE MOROCHO-ZURITA[9], (Member, IEEE), JAVIER MOYA-HONDUVILLA[10], MARÍA T. MANRIQUE-SANCHO[11], MARCELA E. MONTIVEROS[6], ROCÍO NARVÁEZ-BENALCÁZAR[8], JOSÉ DE JESÚS PÉREZ-ALCÁZAR[12], YURI RESNICHENKO[13], AND DIEGO SECO[14]

[1]Facultad de Ingeniería, Universidad ORT Uruguay, Montevideo 11100, Uruguay
[2]Departamento de Geografía, Facultad de Arquitectura, Urbanismo y Geografía, Universidad de Concepción, Barros Arana 4030000, Chile
[3]Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, San Fernando del Valle de Catamarca, Catamarca K4700BIN, Argentina
[4]Centro de Investigación en Ciencias de la Información Geoespacial (CentroGeo),Ciudad de México, 14240, Mexico
[5]Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo 11200, Uruguay
[6]Laboratorio LatinGEO, Universidad Nacional de Catamarca, San Fernando del Valle de Catamarca K4700BIN, Argentina
[7]Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló de la Plana, Spain
[8]Instituto Geográfico Militar, Quito 170102, Ecuador
[9]Facultad de Ingeniería, Universidad de Cuenca, Cuenca 010203, Ecuador
[10]EnvJoy Nature SL, 08018 Barcelona, Spain
[11]Vector ITC Group, 28231 Las Rozas de Madrid, Madrid, Spain
[12]EACH, Universidade de São Paulo, São Paulo 03828-000, Brazil
[13]Facultad de Ciencias, Universidad de la República, Montevideo 11400, Uruguay
[14]Facultad de Ingeniería, IFMD, Universidad de Concepción, Concepcion 3349001, Chile

Corresponding author: Daniela Moctezuma (dmoctezuma@centrogeo.edu.mx)

This work was supported by the IDEAIS Project-CYTED: Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo, under Grant 519RT0579.

ABSTRACT To know the overall regional or international scientific production is of vital importance to many areas of knowledge. Nevertheless, in interdisciplinary areas such as Geographic Information Science (GISc) it is not enough to just count papers published in specific journals. Most of them, as is the case of the International Journal of Remote Sensing (IJRS), welcome GISc papers but are not exclusive to that area so the production assignable to authors in the region must consider not only affiliation but also whether or not each paper falls into the theme of GISc. IJRS publishes far more papers than any other GISc journal, so it is important to assess quantitatively how many of them are of GISc. In this work, a representative sample of IJRS articles published over a period of almost 30 years was analyzed using a specific GISc definition. With these data, a manual classification methodology through a set of experts was carried out, and a dataset was built, analyzed, and statistically tested. As a result we estimate that between 47 and 76% of the IJRS articles can be considered from GISc, with a confidence level of 95%. Aside from the primary goal, this set could be used as a gold standard for future classification tasks. It constitutes the first GISc dataset of this kind, that may be used to train artificial intelligence systems capable of performing the same classification automatically and in a massive way. A similar procedure could be applied to other interdisciplinary fields of knowledge as well.

INDEX TERMS Gold standard, manual classification, indexer consistency, geographic information science.

I. INTRODUCTION

One of the priorities of scientific journals is to be indexed in regional or international catalogs, and those of the Geographic Information Science area (GISc henceforth) are no exception. For example, one of the entry requirements in order to be indexed by the Scientific Electronic Library Online - SciELO,[1] is that a journal relevant to the subject area of Earth Science should publish at least 45 articles per

The associate editor coordinating the review of this manuscript and approving it for publication was John Xun Yang.

[1]A catalog system of scientific journals on the Internet that is very popular in Latin America

year. Candidate journals then have difficulty determining whether the gross academic output to which they potentially have access reaches the minimum threshold. Although this bibliographic exercise seems *a priory* straightforward, this is not the case for GISc. It is not a homogeneous and strictly defined discipline, and that there is no consensus among GISc researchers on the relevant publication outlets [1]. Even for those where ample agreement exists, Caron *et al.* [2] state that it cannot be assumed that all articles published in those journals belong to the GISc area. They are valid outlets for GISc authors, but they are not exclusive. The name of the journal itself generally provides a first hint to the target scientific or technological field. However, as a discipline is usually broad, composed of sub-disciplines and is influenced by other disciplines or areas of knowledge, a journal usually covers a non-exhaustive number of topics of interest that can expand the core discipline or area of knowledge.

The GISc term was originally coined by Michael Goodchild [3], who from the beginning admitted the difficulty of defining and finding a consensus definition. He stated that ''[. . .] GISc is the science behind the systems, in other words the scientific knowledge on which the GIS (Geographic Information System) is based'' [4]. Since then, some other authors (e.g. [5]–[7]) have attempted to specify and narrow the breadth of the original GISc description. For example, the definition of GISc seen as ''[. . .] the set of fundamental issues raised by the technology or the critical issues that arise when the technology is employed. These would include accuracy and uncertainty, scale, and the methods used to capture the infinite complexity of the real world in binary digits'' [6] still remains inclusive and does not provide concise criteria to uniquely label a work as GISc.

The fact that there exist numerous broad and inclusive interpretations and descriptions [8], [9], without yet reaching a consensus on a global, widely-accepted definition of GISc is an impediment to unambiguously determine whether or not an article belongs to GISc. In words of Held, Laudel and Gläser [10], lacking an accepted definition of GISc is equivalent to lacking a ''fundamental truth'' on which to build the identity of that science. Indeed, the authors already warn about the diffuseness and vagueness of scientific areas that do not have ''natural'' boundaries.

Recently, the authors in [11] addressed these issues and proposed a new operational definition driven by a set of actions (verbs), which are related to the science in question, making it easier to know if an article deals with or covers any of such actions. Using such a definition as a starting point, the contributions of the presented work are two-fold. The first is an estimate of the proportion of papers published in a given journal that (according to such definition) can be labeled as GISc. Under mild assumptions, we can later apply this rate to the total number of works with Latin American authors. The second contribution is a bibliographic metadata dataset that has been manually labeled (GISc, NoGISc) by a group of experts. Since there is no true or false classification in cases like this, most of the standard classification accuracy metrics could not be applied. In order to be considered a reliable dataset, we computed the inter-rater reliability thus following best practices. Datasets like this are known as gold standard, and they would be an essential input in the development of a future intelligent system to automatically classify new articles as GISc or NoGISc.

The rest of the article is organized as follows: after the Introduction Section, the background and previous work of the problem is analyzed in Section II, both in the classification of journals and individual articles. It will also justify the need to generate a gold standard, essential to train future artificial intelligence systems to tackle this problem automatically. Next, in the Materials and Methods section, aspects of the process are analyzed, from the definition adopted, to the selection of experts, the sampling criterion used and the design of the questionnaire offered. Aspects of the process, which consisted of two rounds of evaluation of 208 articles carried out by 14 experts, as well as the statistical analysis carried out will be also detailed. In the Results and Discussion section, the results are presented and a brief discussion of them is made. Finally, the Conclusions and future work are presented.

II. PREVIOUS WORK

This section describes the two approaches to do GISc classification: the journal-level and the article-level. The need for a gold standard is also pointed out.

A. GISC CLASSIFICATION AT JOURNAL LEVEL

Some authors have previously attempted to identify the most appropriate journals to publish in the GISc area. For example, Caron *et al.* [2] used the Delphi method to build a list of 46 journals organized in four categories, with IJRS in the first category. Kemp, Kuhn, and Brox in [12] presented a similar methodology to analyze GISc production, classifying the journals in three categories. The first earned a 'leading' rating from most respondents. The second collected those mentioned by a substantial number of respondents, and the third contained the rest. The authors did not generate a strict categorization acknowledging the weaknesses of their procedure. Again, IJRS was classified in the 'leading' category, underlining its importance in the GISc area. Scarletto [13] also built a list of GISc journals but with a different procedure. She selected four grassroots journals and analyzed the source of the citations from 2008-2010. She included 2,070 journals in English, which after further analysis were reclassified into three groups. The first group included 23 journals, the second 190, and the third the remaining journals. Notice that the order of the list represents the certainty of the journal membership in the GISc area. In this case, although the choice of the initial seed of journals was arbitrary, the methodological procedure is objective. The IJRS journal belongs again to the first group. Finally, Biljecki in [1] generated his own list. He pointed out that the lack of a well-established definition of GISc had led to the ranking of journals being somewhat arbitrary. In particular, he proposed removing from the list

those journals whose top priority is not GISc, leaving IJRS out in this case. This is a good example of the necessity to use clear definitions about what is and what is not a work in the GISc research area. In all cases, it must be emphasized that the previous journal lists indicate those who accept GISc articles, but are not necessarily journals exclusive to GISc. This means that not all the works published in any of these journals could be automatically classified as GISc. Therefore, we start from the idea that a methodological alternative is necessary to inform us when an article in a journal is or not from GISc, which can be articulated through a classification at the article level rather than at the journal level.

Shu *et al.* in [14] state that sometimes directly using the classification at the journal level places half the articles in the wrong classes, which will be seen to be consistent with our results (see Section III). In the case of GISc, none of the classifications in vogue (see the list in [15]) clearly distinguishes it from the other sciences. Therefore, relying on these pre-existing listings will be useless in the case of GISc.

B. GISC CLASSIFICATION AT ARTICLE LEVEL

Related literature is scarce when it comes to classifying GISc articles, although there are proposals for viable procedures to do so. Milojević in [16] starts from the hypothesis that journals can belong to one or more categories at the same time, but it is not so frequent that the same happens for an article. To illustrate that, Milojević describes a method for classifying a large volume of articles into the nearly 250 classes provided by Web of Science (WoS in what follows). Some classes included the adjective Multidisciplinary, while the rest were more specific to a particular area (see a list in [16]). His proposed method is as follows: first all journals cataloged with the adjective Multidisciplinary (Multidisciplinary Physics, Multidisciplinary Geosciences, etc.) were rejected while the rest might belong to one or more classes (Biology, Geology, Physics, History, ...). For each WoS category, it is possible to form sets with those journals that belong exclusively to one category. At the article level, the author's hypothesis is that by citing a majority of journal papers from a single class, the work can be assigned to that category. The procedure is iterative, aiming to use in a second stage the categories assigned to the articles that originally did not have one.

Nevertheless, the essential limitation of this procedure is that journals need to be pre-assigned to classes. In the case of GISc, in the listing by Caron *et al.* in [2] there are 46 journals, 17 of which meet the criteria of Milojević [16]. Among them, there is one from Experimental Psychology, one from Geology, two from Remote Sensing and the rest are from Geography. Therefore, applying the Milojević criterion, it will not be possible to infer, from a majority of journals classified as Geography, that the articles will belong to GISc and not to Geography since GISc is not a category foreseen by WoS. Similar results can be derived from the categories from SCOPUS. Starting again with the listing by Caron *et al.* in [2] there are 14 journals with unique SCOPUS category. Five from 'Earth and Planetary Sciences: General Earth and Planetary Sciences'. Three from 'Social Sciences: Geography, Planning and Development'. Two from 'Earth and Planetary Sciences: Earth-Surface Processes', and one from each of 'Engineering: General Engineering', 'Earth and Planetary Sciences: Oceanography', 'Earth and Planetary Sciences: Computers in Earth Sciences' and 'Social Sciences: Social Sciences (miscellaneous)'. It can be stated that not all journals belonging to those categories are specifically related to GISc.

Another alternative is to identify the production in a research area based upon article networks [10]. Nevertheless, despite a promising and interesting approach, conclusive results are still missing.

C. THE NEED OF A GOLD STANDARD

While there are several ways to perform automatic classifications based on citations, references, etc., they all share the problem of the accuracy of their results. Also, most of the prior related works results are based on manual and time-consuming tasks. This leads us to the idea that it is necessary to have at least a partial classification, in which the classes are previously assigned and interpreted, assuming that this classification is correct. For this purpose, it would possible to compare two or more classifications trying to confirm a substantial agreement between them. Klavans and Boyack in [15] comment that there are several automatic classifications, both at the journal and article level, but few have tried to rigorously measure their accuracy. Cabitza *et al.* [17] recently propose (in the case of a manual classification) a procedure that could provide a solution by incorporating the confidence of the experts in each of the answers provided.

Both Waltman and Eck [18], and Klavans and Boyack [15] point out the importance of having a gold standard against which to compare automatic classifications. A gold standard would be a set of articles classified in the classes of interest (GISc in our case) to which high accuracy is attributed. According to the authors, this gold standard could be obtained manually or automatically. But usually the first classification is acquired by human experts as a preliminary step to build intelligent automatic classification systems. Klavans and Boyack [15] identify as members of a possible gold standard those articles with more than 100 references. Through their citations, the authors establish a network of connections that allows them to characterize a field of knowledge. Thus, the articles that cite another pre-classified as GISc would make up a set of articles that would inherit that pre-classification. Nevertheless, characterizing them through the number of their own references is a possible procedure, but it is far from usual.

In a paper devoted to developing a ground truth for a specific field, Held *et al.* [10] mentioned as a first option the use of a gold standard (if available) as a surrogate. Second option was to rely on expert validation, and a third to use already available reliable classifications. To the best of our knowledge, the GISc community lacks a gold standard for discerning whether a work belongs to GISc. Regarding

experts, the authors stated that no individual should provide expert validation alone, even if he/she is an expert in the field under scrutiny. Also, it is very important that any strategy to build a gold standard should be checked against a shared definition of the scientific topic. Nevertheless, this is not the case for GISc.

Hence, this work aims to generate the first gold standard for the case of GISc area. In this case we only include articles published in IJRS. A concrete and operational definition of GISc is used, relying on a set of experts to classify a sample of a reasonable number of cases according to the work of Held, Laudel and Gläser [10].

III. MATERIALS AND METHODS

For our purposes we adapted the procedure HELP [19] designed for labeling students. It has three main steps: pre-labeling, labeling and post-labeling. The first one includes another two: Planning and afterwards Labeler recruitment, Training and Evaluation. We will provide now a summary of the activities and later we will expand on them.

A. PRE-LABELING

1) PLANNING

To assign an article to the category GISc or NoGISc, we used a two stage questionnaire. Details will be provided below, but the first stage requires confirmation/denial that the emphasis of the paper is on methodology. If not, the papers is classified as NoGISc without further analysis. In order to discern the emphasis the procedure pose no special requirement for the experts. The second stage requires evaluating the correspondence of each article with a specific definition, an activity which requires that the expert must be familiar with the terminology. Thus, the required professional characteristic is to actively work with Geographic Information, either in the academia, government or industry.

A key aspect for the success of the classification process was the selection of a definition of GISc. After evaluating the alternatives available in the literature, the one recently proposed by López-Vázquez *et al.* in [11] was used:

GISc definition: "Geographic Information Science is a formal science that studies the methods to capture, store, analyze, model, represent, exchange and manage N-dimensional spatial data".

A significant amount of time was devoted to preparing and designing the forms. We tested and refined the procedure with just three experts, working with IJRS papers as well as from other journals also included in the list of Caron *et al.* [2] (namely 'Cartographica: The International Journal for Geographic Information and Geovisualization';'International Journal of Geographic Information Science' and 'Journal of Spatial Science'). Assuming that the order of the papers is random, we used systematic sample to build a meaningful chunk of data ([19]) choosing for example the paper located fourth in the issue #3 of the last 20 years. Its results are not reported here, but the refining process of the forms was driven to improve the reliability index (to be defined below) among the answers of the three experts. Afterwards, a video explaining the procedure was prepared.[2]

TABLE 1. Description of the experts' profiles. CS denotes computer science.

Gender	Academic	Training	workplace	h-Index (Scopus)	h-Index (Google)
F	MSc	Geo	Government	NA	3
F	MSc	Geo	Government	NA	NA
M	PhD	Geo	Company	NA	4
M	MSc	Geo	University	1	3
M	PhD	CS	University	10	14
M	MSc	Geo	Government	NA	NA
M	PhD	CS	University	NA	7
M	PhD	CS	University	5	11
F	MSc	CS	University	NA	NA
F	PhD	CS	University	9	13
M	PhD	CS	University	18	27
M	PhD	Geo	University	6	7
F	MSc	Geo	Government	NA	NA
F	PhD	Geo	Company	NA	3

2) LABELER RECRUITMENT, TRAINING AND EVALUATION

Using Purposive Sampling [20], a group of 41 GISc experts was initially invited, 32 of whom held a PhD level. Of those who accepted, a group of 14 experts was finally selected. The majority (9/14) were primarily doing research (including Geoinformatics, Geomatics, Geography, Semantics, Artificial Intelligence, Machine Learning, Geodata, SDI, Visualization, Language Processing, Usability, etc.). Table 1 summarizes the individual profiles of the group of experts. It is worth noting the diversity of the expert group, regarding workplace and discipline. The corresponding h-Index range (as of the time of writing) is included as an indicator of their research expertise in the field.

It should be clear that each of the experts had experience in just certain aspects of this scientific field. For that reason, it was assumed that while the sum of these individual experiences could cover a substantial portion of the GISc area, it was unlikely that the group of experts as a whole would be fully competent in the entire field [10].

As a universe of interest, the nearly 10,000 articles that have appeared in IJRS since Goodchild in 1992 [3] first mentioned the term Geographic Information Science were potentially considered. For our study, this journal has several characteristics that make it interesting:

- It does not define itself as a specialized journal in GISc, so it is reasonable to try to estimate the proportion of GISc articles published therein.
- As indicated in López-Vázquez and Bernabé-Poveda in [21], IJRS has an important weight in the production of the Latin American region (482 articles out of a total of 2008, in the period 2009-2019).

The universe of articles to be considered was obtained after querying SCOPUS, covering the period 1992-2020 (up to volume 41, #13). To consider changes in editorial priorities that may have occurred over time, a stratified sampling [20] was considered.

[2]https://tinyurl.com/3yp4x4nc

In order to evaluate agreement levels among the experts, the 14 selected reviewers were provided with the same set of 40 IJRS articles. This corresponds to the first part of a scheme called fully crossed design [22] in order to keep the revision effort limited. These 40 articles were drawn at random from a set of 58 also constructed at random at the rate of two articles for each of the 29 years under analysis (1992-2020). The overall process is sometimes denoted as Two-Stage sampling [20]. The number 40 was selected to keep the evaluation effort under control. To assess agreement between word count responses (required by the query about emphasis), the Krippendorff's alpha [23] was used for ratio-type data (quantitative). Although the first calculated statistic is not directly related to the outcome of interest, it allowed generating confidence in the procedure. To evaluate the binary GISc/noGISc classification, we used three metrics. The first one is the so-called Simple Agreement, the second the Krippendorff alpha index [23] but applied to nominal data (qualitative), and the third the Gwet AC_1 index [24]. Any of them assesses the reliability of the final result. Following the criterion of the artistic competitions of the Olympic Games, two experts were identified whose withdrawal made respectively maximum and minimum the Krippendorff alpha of the classification calculated with the remaining 13. The numerical effect turned out to be minimal. The indicator fluctuates only between 0.30 and 0.25, which shows a consistent behavior compared to the rest of the group. After eliminating the two experts, a majority voting scheme among the 12 experts was used for the classification.

B. LABELING

Once the reliability values among the experts were obtained, a second set was built also by stratified sampling. Following the indications of Hallgren [22], a set of 168 articles was randomly selected, also taken from a random sample of 6 per year over the period under study. Each of the 12 experts received a personalized list of 42 articles systematically selected from the set of 168, assuring that each article receives exactly three evaluations from different experts. To preserve anonymity, the two omitted experts each received also a block of 42 articles. The resulting GISc/NoGISc classification of the 168 articles was calculated by majority vote of three opinions. Unlike suggested by [19] we have not formally requested comments about the procedure but informal ones have been received and considered.

C. POST-LABELING

The answers from the experts were now three per paper. We computed again the same reliability indices. Because the answer is just a binary one there is no chance to detect and remove outliers, as suggested by [19]. Thus, the final class will be decided again by majority voting. The numerical results will be presented below.

D. FURTHER DETAILS ABOUT QUESTIONNAIRE DESIGN

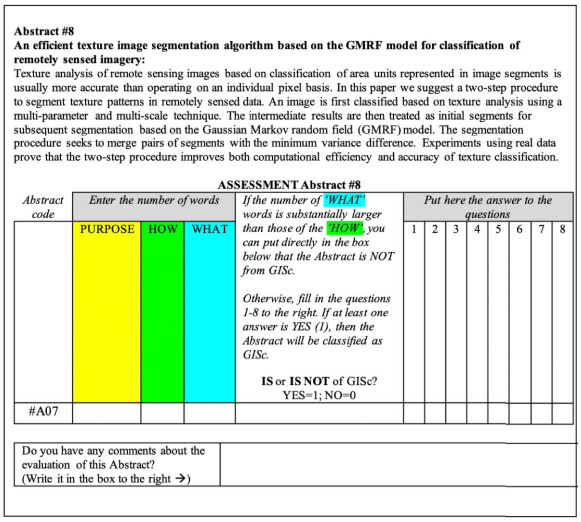
Consistency among rater responses was a primary concern, and therefore efforts were made to measure it. Following [19] the inter-coder reliability was used as a metric of evaluation. According to Lacy *et al.* in [25] the results in terms of reliability measure the appropriateness of the questionnaire more than the competencies of the experts. Due to the peculiarities of GISc the procedure to follow required first to discern the emphasis of the paper. Every sentence of the abstract will thus be classified in three mutually excluding categories denoted by colors and afterwards the number of words involved in each colored part will be counted. Depending on the results the paper will be classified as NoGISc and the evaluation ends, or a second stage will be performed. This criterion was in agreement with Scheider *et al.* [8] who pointed out as characteristics of the GISc theme the emphasis on the HOW and PURPOSE. In his words, the HOW describes the methodology used to solve a problem and the PURPOSE shows the reason for its practical application. In cases where there was no novel methodological contribution, the work is not considered as GISc but rather described the use of GISc for the benefit of empirical science [8]. Otherwise, the expert considered some questions associated with the definition of GISc which led directly to the requested labeling.

The three exhaustive categories (PURPOSE-HOW-WHAT) used to classify the sentences of an abstract were a simple but effective mechanism to find out in which part the author had put more emphasis. The definitions of these parts-of-abstract are as follows:

- PURPOSE (motivation and objective of the work). Explain the problem to tackle. It is not related to the research method, and is usually a simple description of the problem addressed.
- HOW (methodology or procedure followed). The methodological novelty must be described in this part.
- WHAT (description of the result or product). The result is generally associated with the requirements of some empirical evidence.

In the absence of better criteria, it was decided that an indirect way to measure the emphasis that the author places on each of the three blocks is to count the words involved in the abstract of the article. Taking into consideration the above, the final decision was made based on the following assumptions:

- If an article places considerably emphasis on describing the WHAT (results), it can be assumed that the article is about the application of an already known methodology to a field of empirical science and the HOW (the methodology) has less interest. It is even possible that a variant of a well-known methodology is applied, but clearly the author does not consider it to be the most important aspect in the article because it was not reflected in the abstract. Therefore, it was concluded without further considerations that the article was not about GISc but about an empirical science that uses GI to obtain its results.
- If HOW has a greater emphasis or at least comparable to WHAT, then the experts proceed to a second stage of the questionnaire in which they compare the content of the



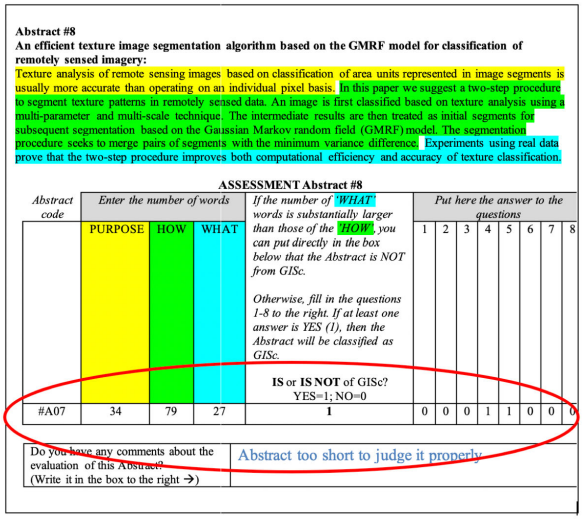
Abstract #8
An efficient texture image segmentation algorithm based on the GMRF model for classification of remotely sensed imagery:
Texture analysis of remote sensing images based on classification of area units represented in image segments is usually more accurate than operating on an individual pixel basis. In this paper we suggest a two-step procedure to segment texture patterns in remotely sensed data. An image is first classified based on texture analysis using a multi-parameter and multi-scale technique. The intermediate results are then treated as initial segments for subsequent segmentation based on the Gaussian Markov random field (GMRF) model. The segmentation procedure seeks to merge pairs of segments with the minimum variance difference. Experiments using real data prove that the two-step procedure improves both computational efficiency and accuracy of texture classification.

ASSESSMENT Abstract #8

Abstract code	*Enter the number of words*			*If the number of 'WHAT' words is substantially larger than those of the 'HOW', you can put directly in the box below that the Abstract is NOT from GISc.* *Otherwise, fill in the questions 1-8 to the right. If at least one answer is YES (1), then the Abstract will be classified as GISc.* **IS** or **IS NOT** of GISc? YES=1; NO=0	*Put here the answer to the questions*							
	PURPOSE	HOW	WHAT		1	2	3	4	5	6	7	8
#A07												

Do you have any comments about the evaluation of this Abstract? (Write it in the box to the right →)	

FIGURE 1. **Visual aspect of one of the summaries received by the experts, with the corresponding empty evaluation table.**



Abstract #8
An efficient texture image segmentation algorithm based on the GMRF model for classification of remotely sensed imagery:
Texture analysis of remote sensing images based on classification of area units represented in image segments is usually more accurate than operating on an individual pixel basis. In this paper we suggest a two-step procedure to segment texture patterns in remotely sensed data. An image is first classified based on texture analysis using a multi-parameter and multi-scale technique. The intermediate results are then treated as initial segments for subsequent segmentation based on the Gaussian Markov random field (GMRF) model. The segmentation procedure seeks to merge pairs of segments with the minimum variance difference. Experiments using real data prove that the two-step procedure improves both computational efficiency and accuracy of texture classification.

ASSESSMENT Abstract #8

Abstract code	*Enter the number of words*			*If the number of 'WHAT' words is substantially larger than those of the 'HOW', you can put directly in the box below that the Abstract is NOT from GISc.* *Otherwise, fill in the questions 1-8 to the right. If at least one answer is YES (1), then the Abstract will be classified as GISc.* **IS** or **IS NOT** of GISc? YES=1; NO=0	*Put here the answer to the questions*							
	PURPOSE	HOW	WHAT		1	2	3	4	5	6	7	8
#A07	34	79	27	1	0	0	0	1	1	0	0	0

Do you have any comments about the evaluation of this Abstract? (Write it in the box to the right →)	Abstract too short to judge it properly

FIGURE 2. **The result of the evaluation of the article. Note the coloring in the text, and the inserted numeric values. In this case there were also comments.**

Article Summary with the proposed definition to finally decide whether or not the article is about GISc.

- In our proposal, the space dedicated to PURPOSE is not considered in the decision.

For an easier visualization, it was proposed to the experts to mark each of the parts of an abstract (PURPOSE, HOW and WHAT) in a different color (see Figure 1 as example).

After that, the experts must do the following:

- Count the number of words in the sentences according to each color (see Figure 2);
- If the number of words of the WHAT section is significantly greater than that of the HOW section, the article is marked as ''NOT belonging to GISc'' and the evaluation ends.
- If the number of words of the HOW section is significantly greater than that of the WHAT section, the expert proceeds to answer eight additional questions (see Figure 3). Since the GISc definition provided is action-based (capture, store, analyze, model, represent, exchange, manage), the expert checks whether the abstract describes or makes reference to any of these verbs. If any of the eight questions received an affirmative answer, the article is then marked as GISc.
- In the case of a similar number of words, the expert should apply their own criteria. There is an optional text field for comments (see the bottom of Figures 1 and 2).

	The "HOW" of the article is mainly about ...	Yes=1 No=0
1	Describe processes and procedures for capturing spatial data Capture specifications. Sampling strategies; Use of volunteers; Sensors and Sensor Networks, etc.	
2	Define aspects of spatial data storage Database structure; Integrity constraints, etc.	
3	Aspects related to semantics Checklists; Gazetteers; Thesauri; Ontologies, etc.	
4	Treatment of raw spatial data (input and output data is of the same type) Fusion; Integration; Filtered out; Data cleansing; Estimation of quality parameters; Comparison between several methods for the previous or improvement of an existing one; Data format transformation (Data munging or data wranglig); Analysis of data heterogeneity etc.	
5	Analyze or mathematically model spatial data (data enters and new data comes out) ATTENTION: It must be a sufficiently general treatment, not tied to a specific problem Error and Uncertainty propagation; Classification methods; Pattern recognition; Model calibration; Simulation; Generalization, etc.	
6	Human-machine interaction to communicate spatial data Cognitive aspects; Geovisualization; Usability; Augmented reality and virtual reality etc.	
7	Machine-to-machine sharing and dissemination of spatial data Interoperability; Exchange standards, Cloud computing etc.	
8	Relationship of spatial data with society Data governance; Digital government; Spatial Data Infrastructure; Legal, economic and security aspects of access; Integrity and authenticity; Privacy, etc.	

FIGURE 3. **Questions to be answered by experts if the emphasis of the article is on HOW.**

E. FURTHER COMMENTS ABOUT THE STATISTICAL ANALYSIS

1) MEASUREMENT OF THE QUALITY OF THE RESULT

In the area of remote sensing, the results of a classification are usually evaluated using indicators of the Cohen kappa type [26], comparing against the reference values obtained in the field or from an independent source of greater accuracy. The same kappa indicator could also be used using two independent classifications, without involving reference data. Although there are serious and well documented concerns about it (see [27]–[30]), the Cohen kappa index is by far the most widely used to characterize the quality of a classification. Others also derived from the Confusion Matrix are Validity, Reliability, Sensitivity, and Specificity. In the case under analysis in this work, none of these metrics can be used for two reasons. First, there is no reference classification or ground truth to compare to. Second, because the set of articles is not evaluated by two but by several experts. Therefore, the quality must be evaluated with another criterion.

According to some authors (see for instance [22] and [31]), in the expert's responses there are two magnitudes of interest: validity and reliability. The latter can be between the expert and himself (intra-rater reliability) or between several experts (inter-rater reliability). To some extent these magnitudes are analogous to accuracy and precision, common terms in many sciences. Accuracy indicates or measures how close the result

is to its true value (which is always inaccessible), while precision measures the mutual agreement between the measurements, a quantity that does not require access to the true value. In addition to the terms mentioned there are others such as Intercoder reliability and even Intercoder agreement that Zhao *et al.* in [32] consider equivalent to Inter-rater reliability although this opinion is not entirely unanimous (Lovejoy *et al.* [33]). In any case this parameter can be assessed in various ways. The literature recommends jointly considering at least three indicators ([33]–[35]). To follow this suggestion and ensure an adequate evaluation, in this work three metrics were used, namely:

- The Simple Agreement is sometimes also called the Holsti reliability coefficient [36] and consists of dividing the total of cases with agreement between the experts by the total of evaluated cases. The ratio oscillates in the interval [0,1]. Although it has many limitations, there are authors who recommend its use when there is not a marked imbalance between the classes.
- Krippendorff's alpha is a measure of agreement that takes values in the interval [−1,1] (as the well-known Cohen's kappa does), with case 1.0 corresponding to perfect agreement, 0.0 to a random classification and −1.0 to a perfect disagreement (inverse agreement). Delgado and Tibau [29] show that in certain circumstances Krippendorff's alpha and Cohen's kappa coincide. Although Krippendorff alpha has been used in the field of remote sensing (see Rosenfield and Fitzpatrick-Lins [37] and Kerr *et* . [38]), it is also the standard indicator in other areas, for instance communication content research [33].
- Like alpha and kappa, AC_1 is also an indicator that corrects for random coincidences. Unlike the previous ones, AC_1 assigns a different behavior to the experts depending on whether the article is 'easy' or 'hard' to classify. If it is Easy, the experts will presumably classify it correctly. If is Hard, it is assumed that they will decide at random with uniform distribution among the available options (in this case, GISc/NoGISc). The AC_1 is also in the interval [−1,1].

There is some disagreement on the acceptable or adequate values that these statistics would take. In the case of Cohen's kappa, Landis and Koch in [39] suggested without much justification cut-off thresholds to qualify adjustments from poor to near perfect, thresholds that have been frequently cited in the literature. Table 2 lists these values, which will be mentioned later.

Bornmann *et al.* in [40] recognize that, in practice, there are frequent cases in which the reliability indices do not reach the high values specified by Landis and Koch. This is also expressed by Lovejoy *et al.* [33] who points out as a problem the lack of concrete rules and not merely speculation about the values that should be acceptable. In fact, Lacy *et al.* [25] go further and question the specification of values to be achieved. The authors point out that these numerical indicators are useful during the protocol adjustment process, giving objective criteria to improve its performance. These improvements in the protocol will allow the values to be increased somewhat, always within what the problem makes possible. There are works such as that of Thompson and Walter in [41] or that of Zapf *et al.* [31] that indicate that the thresholds suggested by Landis and Koch cannot always be assumed to be achievable since they depend (among other things) on the prevalence of the defined categories.

TABLE 2. Threshold values commonly used to qualify the results. Taken from Landis and Koch [39].

kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

In addition, these indices are not infallible, and there are well-known situations (known as 'kappa paradoxes', see Cicchetti and Feinstein [42] in which, despite a high agreement, the index takes a low or very low value (Feinstein and Cicchetti [43], Gwet [24]). A similar situation occurs for other statistics [34]. Given the relationship of kappa with Krippendorff's alpha, it can be assumed that the threshold values proposed for kappa would also serve as guidelines to alpha. Alternatively, the values suggested by Krippendorff could be used.

2) ESTIMATION OF THE PROPORTION OF GISC ARTICLES IN IJRS

Considering the total sum of 208 (40 + 168) classified articles, the proportion of GISc articles in the sample can be initially established. However, in addition to the value that can be obtained with this sample, it is interesting to estimate its confidence interval applying the re-sampling technique called Bootstrap proposed by Efron in [44], widely used in experimental sciences. It consists of the random extraction with replacement of a total of 208 articles, where the 208 articles are available to be chosen. It is thus clearly possible that some article is repeated, even more than once. Each time a set of 208 articles is assembled (called a re-sample) the percentage of cases classified as GISc can be easily calculated. If the operation is repeated many times, a large set of values is built, each of which is representative of the variable of interest for this analysis. The 2.5% and 97.5% percentiles of this population define a 95% confidence interval for it, which is what will be reported in the results. The bootstrap technique is non-parametric, and therefore the re-samples are not expected to belong to any particular probability distribution.

In this sense, all the calculations were carried out in Octave 5.2,[3] and for the inter coder reliability index the toolbox due to Girard[4] was used.

[3]https://www.gnu.org/software/octave/index.html
[4]https://tinyurl.com/2z6dbem4

TABLE 3. Krippendorff alpha values as a reliability indicator for the number of words per class.

	PURPOSE	HOW	WHAT
English (4 reviewers)	0.60	0.67	0.81
Spanish (10 reviewers)	0.55	0.55	0.62

IV. RESULTS AND DISCUSSION

The proportion of GISc-type articles in the set of 208 articles was of 62.5%. To provide confidence levels, the set of 208 papers was re-sampled 10^6 times, and the proportion of GISc articles was calculated for each re-sample. It was found that the re-sampled proportions do not follow a normal distribution. With the population of the proportions obtained, a 95% confidence interval was estimated, which turned out to be [47% 76%].

The length of the interval is relatively large, but for the purposes of this paper it is not a problem. After quantifying the academic production in GISc in the region, the data will be used in support of political decisions for investments in research, offers of specific postgraduate courses, editorial decision-making to apply or not to a re-categorization of one journal, etc. all of which require other inputs as well with perhaps even greater uncertainties. If any of the envisioned applications require smaller uncertainties, the sample sizes will have to be expanded.

In addition to the proportion of GISc articles in IJRS, something should be said about the quality of the data used. The classification in GISc/NoGISc was done in two stages. In the first one, the experts were asked to classify the sentences that appeared in the abstract into one of three categories (PURPOSE, HOW, WHAT). Then the words were counted and if the total of words in the HOW category was not clearly greater than the total in the PURPOSE category then the task was finished and the article was classified as NoGISc. Before continuing, it is important to assess the agreement between the experts even at this intermediate stage. Table 3 shows the results of the first round of 40 articles, for which the classifications of 14 reviewers were available. The reliability index (Krippendorff's alpha in this case) is somewhat lower for the classification in Spanish than in English, but the differences are not substantial. Among the ones available, Krippendorff's alpha is the only reliability index suitable for quantitative answers. The agreement can be classified as Moderate or Substantial according to the classes of Landis and Koch (1977). The relatively high level of agreement among the experts is a positive and surprising result, as it empirically corroborates that the instrument led to consistent results in terms of (PURPOSE, HOW, WHAT). Indirectly, this show that it is capable of detecting the emphasis of the article.

To the extent that the word count in each class could discard an article for the rest of the process, the good agreement achieved is considered significant, regardless of the language in which the evaluation was carried out.

After this verification, an Inter Rater Reliability analysis was carried out to characterize the level of agreement that the experts had when making their classification as GISc/NoGISc. It should be noted that the prevalence of the GISc class compared to the NoGISc class was close to 62.5%, which removes the problems of the indices associated with imbalances between classes. In the first stage and after discarding two experts with the mentioned criterion of artistic gymnastics, the 40 articles received 12 evaluations each, being 0.67, 0.34 and 0.36 the values of the reliability index considered (Simple Agreement, Krippendorff alpha, and AC_1 respectively) as shown in Table 4. According to the usual criteria, the adjustment can be described as Fair. When repeating the calculations for the set of 208 articles, the results were very consistent with the previous ones, reaching 0.67, 0.34 and 0.35, respectively. These results validate the option followed by using only three experts per article for the last 168 works, which allowed, with the same number of experts, to generate a substantially larger set of classified articles. In a very recent work, and according to Cabitza *et al.* [17], if the results of the Inter Rater Reliability are similar with 3 and with 12 experts, this could be interpreted as that the (unknown) accuracy exceeds 85%. This is consistent with what was observed when deciding which experts to remove from the set of 14, for which all the possibilities of using 13 of them were evaluated. An indicator of the high competence of the experts is the low variation of the Krippendorff alpha in the 14 alternatives, which ranged between 0.25 and 0.30.

TABLE 4. Results of the inter rater agreement in the classification of articles as GISc/NoGISc.

		Holsti	alpha	AC_1
40 papers, 12 opinions/paper	mean value	0.67	0.34	0.36
	95% interval	[0.64 0.71]	[0.25 0.42]	[0.29 0.43]
168 papers, 3 opinions/paper	mean value	0.67	0.34	0.34
	95% interval	[0.55 0.79]	[0.11 0.57]	[0.09 0.60]
208 papers	mean value	0.67	0.34	0.35
	95% interval	[0.57 0.77]	[0.22 0.46]	[0.14 0.55]

The single value of an index may say nothing of its own uncertainty. Following the suggestion of Gwet [24], the variance of each estimator was estimated with the Jacknife method, then applying a variance correction factor. The confidence intervals assuming normality turn out to be relatively wide, which is illustrated in Table 4. They are smaller in the case of the first 40 articles, surely reflecting the greater number of opinions available per article. The Krippendorff alpha index shows somewhat greater variability than the other two.

About the results, we could state that:

- The relatively high degree of agreement obtained by the experts in the identification of PURPOSE-HOW-WHAT phrases in the abstract demonstrate that it is a consistent method to identify the emphasis of an article.
- The results confirm that it is not necessary to use a large number of experts per article, since a comparable inter rater agreement was achieved with either 3 or 12.
- The evaluation was a heavy task. The typical time required to read the 40 abstracts, classify into PURPOSE-HOW-WHAT blocks, count the words in each one, and finally compare the abstract with the definition, can be estimated as a few hours.

- If the work with experts is to be repeated in the future, it would be convenient (following Cabitza *et al.* [17]) to ask about the confidence that they themselves attribute to their opinion in each article. According to these authors, this would enable additional quantitative analysis.
- Reducing the length of the Confidence Interval of the proportion of GISc articles will require an increase in the number of articles to be analyzed.
- Our results in the form of a 'gold standard' is valid for articles already published but may become obsolete as the technologies and applications of GISc vary over time. The method will be applicable as long as the definition remains acceptable. The data that support the findings of this study are openly available in https://figshare.com/s/32aa4615e6eb0959c36e.

V. CONCLUSION

The GISc area has a strong multidisciplinary nature, which means that substantive contributions can appear without contradiction in journals devoted to mathematics, geography, computer science, etc. thus making it difficult to quantify the production of the area based upon just journals. Article level classification is required. To tackle this, a sample of papers from IJRS was considered. The existing literature has consistently included and rated IJRS as one of the most important publications in the area, although IJRS does not consider itself as such.

A binary classification was achieved, elaborated by a group of 14 experts applied to a sample set of 208 IJRS articles, which are the result of a stratified sampling of a universe of almost 10,000 articles published in the period 1992-2020. As a result, the estimated proportion of GISc articles in the total published in IJRS for the period considered reaches 62.5%, with [47.5% 76.2%] being the 95% confidence interval. From López-Vázquez and Bernabé-Poveda [21], IJRS holds 24% of the papers published in GISc journals up to 2019 with the participation of authors from the Latin American region. The second journal in importance from the list of Caron *et al.* [2] takes 11%. Assuming that the IJRS articles with coauthors from the region behave like the rest of the population, the production that can be attributed to Latin American authors would be between 10 and 16 articles per year.

The set of 208 scientific articles classified as GISc/NoGISc can be regarded as a 'gold standard' for GISc. This type of datasets is used in the area of Artificial Intelligence, specifically machine learning, to automate classification tasks. Its usefulness thus exceeds IJRS and the period considered. Looking at the future work, with adequate machine learning procedures, it could be possible to generate algorithms capable of replicating the classification and massively apply it to all the scientific literature. This would allow discovering GISc articles published in journals that *a priori* could have been discarded. Or it would also allow the identification and reclassification of journals as specialized in GISc.

All in all, the gold standard dataset could allow the development of systems capable of emulating the work of humans but applying to all the works published in IJRS (instead of a sample) or to another journal in the area. In this way, even admitting that the automatic classification is not conclusive, an estimate of academic production in the GISc area could be obtained and updated regularly, which becomes useful in research policies that involve as diverse activities as academic management decisions, promotion of research areas, launch of new journals, and allocation of research funds.

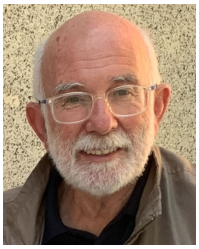
REFERENCES

[1] F. Biljecki, "A scientometric analysis of selected GIScience journals," *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 7, pp. 1302–1335, 2016.
[2] C. Caron, S. Roche, D. Goyer, and A. Jaton, "GIScience journals ranking and evaluation: An international Delphi study," *Trans. GIS*, vol. 12, no. 3, pp. 293–321, Jun. 2008.
[3] M. F. Goodchild, "Geographical information science," *Int. J. Geograph. Inf. Syst.*, vol. 6, no. 1, pp. 31–45, 1992.
[4] D. M. Mark, "Geographic information science: Defining the field," in *Foundations of Geographic Information Science*, vol. 1. London, U.K.: Taylor & Francis, 2003, pp. 3–18.
[5] D. Mark, "Geographic information science: Critical issues in an emerging cross-disciplinary research domain," *J. Urban Regional Inf. Syst. Assoc.*, vol. 12, pp. 45–54, Jan. 2000.
[6] M. F. Goodchild, "Geographic information systems and science: Today and tomorrow," *Ann. GIS*, vol. 15, no. 1, pp. 3–9, Nov. 2009.
[7] M. Duckham, *Geographic Information Science*. Atlanta, GA, USA: American Cancer Society, 2017, pp. 1–13.
[8] T. Blaschke and H. Merschdorf, "Geographic information science as a multidisciplinary and multiparadigmatic field," *Cartogr. Geograph. Inf. Sci.*, vol. 41, no. 3, pp. 196–213, May 2014.
[9] H. Couclelis, "Climbing on a milestone for a better view: Goodchild's 'geographical information science' paper as vantage point and ground for reflection," *Int. J. Geograph. Inf. Sci.*, vol. 26, no. 12, pp. 2291–2300, Dec. 2012.
[10] M. Held, G. Laudel, and J. Gläser, "Challenges to the validity of topic reconstruction," *Scientometrics*, vol. 126, no. 5, pp. 4511–4536, May 2021.
[11] C. López-Vázquez, M. E. Gonzalez-Campos, and M. N. Bernabe-Poveda, "Further steps against the scientific gerrymandering: A new definition of geographic information science," Working Paper, Tech. Rep., 2021.
[12] K. Kemp, W. Kuhn, and C. Brox, "Results of a survey to rate GIScience publication outlets," in *Proc. AGILE*, 2013, pp. 1–8.
[13] E. A. Scarletto, "Mapping the literature of GIS," *College Res. Libraries*, vol. 75, no. 2, pp. 179–201, Mar. 2014.
[14] F. Shu, C.-A. Julien, L. Zhang, J. Qiu, J. Zhang, and V. Larivière, "Comparing journal and paper level classifications of science," *J. Informetrics*, vol. 13, no. 1, pp. 202–225, Feb. 2019.
[15] R. Klavans and K. W. Boyack, "Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?" *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, pp. 984–998, Apr. 2017.
[16] S. Milojević, "Practical method to reclassify web of science articles into unique subject categories and broad disciplines," *Quant. Sci. Stud.*, vol. 1, no. 1, pp. 183–206, Feb. 2020.
[17] F. Cabitza, A. Campagner, D. Albano, A. Aliprandi, A. Bruno, V. Chianca, A. Corazza, F. Di Pietto, A. Gambino, S. Gitto, C. Messina, D. Orlandi, L. Pedone, M. Zappia, and L. M. Sconfienza, "The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability," *Appl. Sci.*, vol. 10, no. 11, p. 4014, Jun. 2020.
[18] L. Waltman and N. J. van Eck, "A new methodology for constructing a publication-level classification system of science," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2378–2392, Dec. 2012.
[19] S. Aslan, S. E. Mete, E. Okur, E. Oktay, N. Alyuz, U. E. Genc, D. Stanhill, and A. A. Esme, "Human expert labeling process (HELP): Towards a reliable higher-order user state labeling process and tool to assess student engagement," *Educ. Technol.*, vol. 57, no. 1, pp. 53–59, 2017.
[20] J. R. Fraenkel, N. E. Wallen, and H. H. Hyun, *How to Design and Evaluate Research in Education*. New York, NY, USA: McGraw-Hill, 2012.
[21] C. López-Vázquez and M. A. Bernabé-Poveda, "La situación de la producción científica latinoamericana en el área de la ciencia de información geográfica," *Revista Cartográfica*, no. 100, pp. 173–193, 2020.

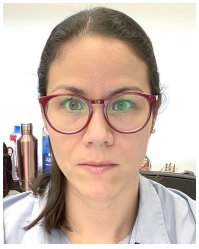
[22] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tuts. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, Feb. 2012.
[23] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educ. Psychol. Meas.*, vol. 30, no. 1, pp. 61–70, 1970.
[24] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *Brit. J. Math. Statist. Psychol.*, vol. 61, no. 1, pp. 29–48, 2008.
[25] S. Lacy, B. R. Watson, D. Riffe, and J. Lovejoy, "Issues and best practices in content analysis," *J. Mass Commun. Quart.*, vol. 92, no. 4, pp. 791–811, Dec. 2015.
[26] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[27] R. G. Pontius, Jr., and M. Millones, "Death to Kappa and to some of my previous work: A better alternative," *Int. J. Remote Sens.*, vol. 32, no. 15, pp. 4407–4429, 2011.
[28] S. V. Stehman and G. M. Foody, "Key issues in rigorous accuracy assessment of land cover products," *Remote Sens. Environ.*, vol. 231, Sep. 2019, Art. no. 111199.
[29] R. Delgado and X.-A. Tibau, "Why Cohen's Kappa should be avoided as performance measure in classification," *PLoS ONE*, vol. 14, no. 9, Sep. 2019, Art. no. e0222916.
[30] G. M. Foody, "Explaining the unsuitability of the Kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification," *Remote Sens. Environ.*, vol. 239, Mar. 2020, Art. no. 111630.
[31] A. Zapf, S. Castell, L. Morawietz, and A. Karch, "Measuring inter-rater reliability for nominal data—Which coefficients and confidence intervals are appropriate?" *BMC Med. Res. Methodol.*, vol. 16, no. 1, pp. 1–10, Dec. 2016.
[32] X. Zhao, G. C. Feng, J. S. Liu, and K. Deng, "We agreed to measure agreement—Redefining reliability de-justifies Krippendorff's alpha," *China Media Res.*, vol. 14, no. 2, pp. 1–15, 2018.
[33] J. Lovejoy, B. R. Watson, S. Lacy, and D. Riffe, "Three decades of reliability in communication content analyses: Reporting of reliability statistics and coefficient levels in three top journals," *J. Mass Commun. Quart.*, vol. 93, no. 4, pp. 1135–1159, Dec. 2016.
[34] X. Zhao, J. S. Liu, and K. Deng, "Assumptions behind intercoder reliability indices," *Ann. Int. Commun. Assoc.*, vol. 36, no. 1, pp. 419–480, Jan. 2013.
[35] G. C. Feng, "Intercoder reliability indices: Disuse, misuse, and abuse," *Qual. Quantity*, vol. 48, no. 3, pp. 1803–1815, May 2014.
[36] O. R. Holsti, *Content Analysis for the Social Sciences and Humanities*. Reading, MA, USA: Addison-Wesley, 1969.
[37] G. H. Rosenfield and K. Fitzpatrick-Lins, "A coefficient of agreement as a measure of thematic classification accuracy," *Photogram. Eng. Remote Sens.*, vol. 52, no. 2, pp. 223–227, 1986.
[38] G. H. G. Kerr, C. Fischer, and R. Reulke, "Reliability assessment for remote sensing data: Beyond Cohen's Kappa," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4995–4998.
[39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[40] L. Bornmann, R. Mutz, and H.-D. Daniel, "A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants," *PLoS ONE*, vol. 5, no. 12, Dec. 2010, Art. no. e14331.
[41] W. D. Thompson and S. D. Walter, "A reappraisal of the Kappa coefficient," *J. Clin. Epidemiol.*, vol. 41, no. 10, pp. 949–958, Jan. 1988.
[42] D. V. Cicchetti and A. R. Feinstein, "High agreement but low Kappa: II. Resolving the paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 551–558, Jan. 1990.
[43] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. The problems of two paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543–549, Jan. 1990.
[44] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.

CARLOS LÓPEZ-VÁZQUEZ (Senior Member, IEEE) received the Ph.D. degree in geomatics from the Royal Institute of Technology, Stockholm, Sweden, in 1997. He is currently a Researcher with Universidad ORT Uruguay. His research interests include GIScience and spatial data infrastructures, as well as statistics and numerical analysis.

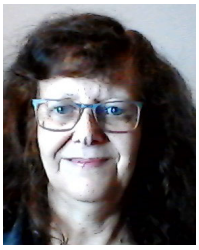
MARÍA ESTER GONZALEZ-CAMPOS received the Ph.D. degree in geographical engineering from the Technical University of Madrid, Spain. She is currently a Professor with the Department of Geography, Faculty of Architecture, Urbanism and Geography, University of Concepción, Chile. She is a Graduate and a Professor of geography with the National University of Patagonia, Argentina. She is also a Specialist in spatial data infrastructure and its implementation with Open-source tools with the Technical University of Madrid.



MIGUEL A. BERNABÉ-POVEDA graduated in surveying and due to his interest in teaching visual communication later graduated in fine arts. He received the Ph.D. degree from the National University of Distance Education in Philosophy and Education Sciences with a thesis on the incorporation of graphic technologies to the cartographic process, in 1994. He is currently pursuing the Ph.D. degree with the National University of Catamarca, Argentina. Throughout his career, he has taught in aspects related to cartography. He is also a Retired Professor and a Researcher with the Technical University of Madrid. His research interests include issues related to the improvement of cartographic writing: graphic semiology, multimedia applied to cartography, communicability, usability, visualization, and all of this applied to spatial data infrastructures.



DANIELA MOCTEZUMA received the Ph.D. degree in computer sciences from Rey Juan Carlos University, Madrid, Spain, in 2013. Since 2014, she has been a Researcher at the Research Center on Geospatial Information Sciences (CentroGEO). Her research interests include machine learning, computer vision, natural language processing, intelligent video surveillance systems, and remote sensing.



ESTHER HOCHSZTAIN received the M.Sc. degree in computer science from the Universidad de Chile, in 1995. Since 1987, she has been a University Teacher and a Researcher at the Universidad de la República, Uruguay. Her research interests include spatial data infrastructures, business intelligence, data/web mining, and statistics.

MARÍA A. BARRERA has been a Researcher at the Geographic Information Technology Laboratory, National University of Catamarca, since 2009. Her research interests include spatial data infrastructure, geospatial data reusability and spatial open data.

CARLOS GRANELL-CANUT graduated in computer engineering, in 2000. He received the Ph.D. degree in computer science from the Universitat Jaume I, Castellón, Spain. He was a Postdoctoral Researcher at the European Commission–Joint Research Centre, Italy. He is currently an Associate Professor at the Universitat Jaume I. His research interests include multi-disciplinary application of geographic information science, spatial analysis and visualization of new forms of spatial data, and reproducibility research practices.

MARÍA F. LEÓN-PAZMIÑO is currently pursuing the Ph.D. degree in geography with the Universidad Nacional del Sur (UNS), Argentina. Since 2005, she has been a Professional at the Military Geographic Institute (IGM), Ecuador. Her research interests include spatial data infrastructures (SDI) being part of the working group responsible for the SDI of the IGM and the Ecuadorian geospatial data infrastructure (IEDG).

PABLO LÓPEZ-RAMÍREZ received the bachelor's degree in physics from the National University, and the M.Sc. degree in geomatics and the Ph.D. degree in geospatial information sciences from the CentroGeo. He is currently an Associate Researcher at CentroGeo. His research interests include construction of geographic databases for the development of urban pollution inventories; the development of socio-technical tools for the construction of institutional data catalogs and geographic knowledge discovery, and data mining.

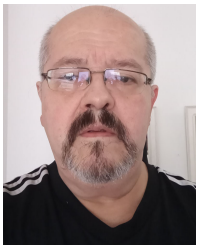
VILLIE MOROCHO-ZURITA (Member, IEEE) received the Ph.D. degree from the Polytechnic University of Catalonia, specialist in spatial data infrastructures (SDIs). He was the Founder and the Director of the Center for Development and Innovation, from 2007 to 2012. He was also the Executive Director of the National Academic Network (CEDIA), from 2009 to 2013. He has been the Director of projects, among others, related to SDIs, since 2008, with the National and International funding. He is currently a full-time Senior Researcher at the University of Cuenca, Ecuador. He belongs to the Department of Computer Science and is the Co-Founder of the Virtual Laboratory of City and Territory.

JAVIER MOYA-HONDUVILLA received the master's degree in geodesy and cartography. He is currently pursuing the Ph.D. degree in geographical engineering specialized in cartography visualization. He is also a Surveying Engineer. He is focusing his research effort on improving the graphic communicability of complex scientific geospatial data. He is also an Independent Consultant in the field of geographic information technologies, graphic semiology, and geospatial data communicability.

MARÍA T. MANRIQUE-SANCHO is currently a Ph.D. Engineer/UX Designer whose specialty sits at the intersection between research, human sciences, technology and innovation. Her passion is to involve users on processes to resolve evidenced-based needs, define strategies and improve users experiences. She has taken part in a wide variety of projects covering areas such as cartography, robotics, education, tourism, marketing, health, food, entertainment, and NGOs.

MARCELA E. MONTIVEROS is currently an Agricultural Engineer and specializes in cadastral and property appraisal. Her research interests include spatial data infrastructures, cadastral management, and GIS.

ROCÍO NARVÁEZ-BENALCÁZAR received the Ph.D. degree in geography. She is currently the Head of the Department of Digital Cartography, Directorate of Appraisals and Cadastre, Municipality of the Metropolitan District of Quito, Ecuador. Her research interests include digital cartography, cadastral analysis, and urban planning.



JOSÉ DE JESÚS PÉREZ-ALCÁZAR received the Ph.D. degree in informatics from the Pontifical Catholic University of Rio de Janeiro, Brazil, in 1996. Since 2014, he has been an Assistant Professor at the University of Sao Paulo. His research interests include semantic web, web of data, data bases, data analytics, and remote sensing.

YURI RESNICHENKO received the M.Sc. degree. Since 2002, he has been working as a University Professor of geographic information technologies with the Faculty of Sciences, University of the Republic, Uruguay. He is currently a Geographer. He has specialized and researched in the production and use of geographic information, particularly related to the spatial data infrastructures.

DIEGO SECO received the bachelor's and Ph.D. degrees in computer science from the University of A Coruña, Spain, in 2006 and 2009, respectively. He is currently an Associate Professor at the Department of Informatic Engineering and Computer Science, Faculty of Engineering, University of Concepción, Chile. His research interests include geographic information retrieval, geographic information systems, and compressed structures and algorithms for textual and geographic data.