International Journal of Geo-Information

MDPI

Article

A Comparison of Cartographic and Toponymic Databases in a Multilingual Environment: A Methodology for Detecting Redundancies Using ETL and GIS Tools

Oihana Mitxelena-Hoyos [1,2] and José-Lázaro Amaro-Mellado [3,4,*]

[1] Instituto Geográfico Nacional-Servicio Regional en Cantabria-País Vasco, 20010 San Sebastian, Spain
[2] Escuela de Ingeniería de Gipuzkoa, UPV-EHU, 20018 San Sebastian, Spain
[3] Departamento de Ingeniería Gráfica, Universidad de Sevilla, 41092 Seville, Spain
[4] Instituto Geográfico Nacional-Servicio Regional en Andalucía, 41013 Seville, Spain
* Correspondence: jamaro@us.es

Abstract: Toponymy, a transversal discipline for geography, linguistics, and history, finds one of its main supports in cartography. Due to exhaustiveness on the territory, cadastral cartography and its toponymy have the ideal characteristics to develop systematic geographical analyses. Moreover, cadastre and geographical names are part of the geographic reference data according to Annex 1 of the INSPIRE directive. This work presents the design, implementation, and application of a methodology based on Geographic Information Systems and Extract, Transform, and Load (ETL) tools for detecting coincidences between the cadastral geoinformation and the official gazetteer corresponding to the province of Gipuzkoa, Spain. Methodologically, this study proposes a solution to the issues raised by bilingualism in the study area. This problem is approached a priori, in the previous data treatment, and a posteriori, applying semantic criteria. The results show a match between the datasets of close to 40%. In this way, the uniqueness and richness of the analyzed source and its outstanding contribution to the potential integration of the official toponymic corpus are evidenced.

Keywords: gazetteer; cartography; cadaster; GIS; tools; toponymy; place name

Citation: Mitxelena-Hoyos, O.; Amaro-Mellado, J.-L. A Comparison of Cartographic and Toponymic Databases in a Multilingual Environment: A Methodology for Detecting Redundancies Using ETL and GIS Tools. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 70. https://doi.org/10.3390/ijgi12020070

Academic Editors: Florian Hruby and Wolfgang Kainz

Received: 1 December 2022
Revised: 5 February 2023
Accepted: 16 February 2023
Published: 18 February 2023

1. Introduction

Landscape, cartography, and toponymy are intertwined because of their links to territories and human activity. All three involve ways of understanding and describing the organization of the space where humans carry out their activities [1–3]. On the one hand, a geographical name has the primary function of naming a place and acting, at the same time, as a semantic and locating function [4,5]. On the other hand, the analysis of the nature of its formation and the determining factors for its survival allow it to become a piece of the landscape archaeology [6], since they bring past realities to today. This patrimonial function of place names can also be observed in a sociolinguistic aspect, especially in regions where contact between languages has given rise to different forms of multilingualism over time [7]. In this context, we must also emphasize that the transversal nature of toponymy makes it necessary to maintain a multidisciplinary perspective in its treatment.

At the same time, cartography is an unbeatable tool for the study, standardization, and dissemination of toponymy, taking advantage of the geographical component of the place name with which the human groups that inhabit, or have inhabited, a territory know and designate it [4]. Cadastral cartography has specific characteristics that place it in a special/peculiar position concerning its toponymic content. According to González García [8], cadastral information is endowed with certain characteristics with respect to the territory due to its status as an exhaustive census record. Since it reflects all the properties on the ground and is continuously maintained, it also facilitates the knowledge of geometric

or other variations in the territory. These characteristics are ideal for combining in studies based on territorial analysis.

Additionally, cadastral information is a repository of geographic and toponymic knowledge. Both the cadastre and geographical names are part of the core (reference data) of the geoinformation of territory, since they appear in Annex I of the INSPIRE Directive 2007/2/EC, "which means that they are considered as reference data, i.e., data that constitute the spatial frame for recognizing geographical location" [9]. Likewise, the Spanish legislation, Law 14/2010 of 5 July, on infrastructure and geographic information services in Spain (LISIGE) establishes that both the Official Gazetteers and cadastral geographic information belong to the Geographic Reference Information (GRI) [10]. This phenomenon is so important that the cadastral parcel has become the working and reference unit for the other GRI features.

Furthermore, the geospatial data-integration capability of Geographic Information Systems (GIS) is beyond doubt. These tools can be used for work in all fields where the location is relevant and are especially useful in multidisciplinary environments [11]. Alternatively, when the volume of data is huge or repetitive geographic processes are to be performed, Extract, Transform, and Load (ETL) tools become extremely useful [12–14].

This work aims to improve the efficiency of integrating toponymic data from different geoinformation sources, using a semi-automatic methodology that minimizes the effect that multilingualism produces in detecting data redundancy. Consequently, this research offers a methodology for the synchronization of different toponymic databases. To this end, we use GIS and ETL tools to integrate place names from official sources of different levels of administration, in the geographical area of the historical territory of Gipuzkoa, Basque Country (*País Vasco* or *Euskadi* or CAPV or CAE), Spain. The databases used have different characteristics in terms of the scale of capture and processing of place names, degree of linguistic standardization of the names, as well as classification into typologies. In this context, there is a possibility of innovation by incorporating foral cadastral cartography in these tasks.

The closest equivalent work, both geographically and conceptually, is that carried out by the Autonomous Community of Andalusia [15], where the integration of cadastral toponymy in its Official Gazetteer (NGA) was addressed with very favorable results [16]. Another close piece of research is the semi-automatic collation work planned for the improvement of the Basic Geographic Gazetteer of Spain (NGBE) [17] called "autocorrection." This gazetteer is handled by the Spanish National Mapping Agency, *Instituto Geográfico Nacional* (IGN). As a result, concordance between elements of the NGBE and the Official Geographic Gazetteer of the Autonomous Community of the Basque Country (NGO-CAE) was found. These two cases serve as a reference to contrast the suitability of the proposed methodology.

The method developed aims to evaluate the complementarity of the datasets to discover their original value and avoid redundancies when working with them jointly. The FME extraction capability has been used in the initial source processing work and the search for combined textual and localization matches. For the design issues of the lexeme-based semantic validation of the results [18], as well as the manual evaluation of the method on a sample of the data, the capabilities of GIS tools have been used. Concerning this research, the study presented here is innovative because it considers and anticipates specific aspects derived from bilingualism, since Basque (or *Euskera*) and Spanish languages are co-official in Gipuzkoa.

2. Related Work

This section presents the relevant publications related to this toponymy investigation from different points of view, such as cartography, GIS, cadaster, and multilingualism. Given that our research is strongly multidisciplinary, some of the cited works could appear (or have appeared) in different subsections.

2.1. On Place Names, Cartography, and GIS

Cartography is a form of representation of a synthesized geographic space, and toponymy identifies its different elements. The place names on a map allow the identification of the geographical elements represented. In addition, as a graphic element of the map, the label itself becomes a vehicle for transmitting information about the element through the visual variables of the field of graphic semiology [19,20]. As geographic information, it is inevitable to conclude that the most efficient way to manage this dataset is the usage of Geographic Information Systems. Therefore, the design and implementation of the GIS project are crucial to respond to current requirements [21].

Regarding the localizing function of toponymy, different works prove this capacity. Thus, Frajer and Fiedor made up a toponymic GIS to find out extinct water bodies [22]; Gordova et al. dealt with geographical names to solve historic-geographical issues and created regional atlases [23]. Other researchers have assessed place names as geographical information tools [24]. Giraut and Houssay-Holzschuch suggested a theoretical framework in order to analyze place naming regarding geopolitics and power relations [25].

On the one hand, the perspective of physical geography has been widely portrayed through place names [26]. On the other hand, different works interrelate the names of different entities that represent types of structuring landscape elements [27], such as roads [28], hydrology [22], and constructions, whether traditional [29] or urban [30]. All this does not undermine the fourth dimension, the temporal aspect present both in the historical testimony [31,32] and in the dynamism of the landscape, reflected in the labeling of maps [33] through the diachronic perspective in historical research [34]. Furthermore, with this temporal component, we cite its localizing capacity [35]. On the other hand, from individual work to distributed and networked systems, the way of accessing and exploiting geographic information, including cadastral information, has radically changed [36].

In addition, ETL assists in the management of different databases by adding power to the analysis. These tools also facilitate the preparation of the data for analysis while providing the possibility for optimizing the databases' structure [37], both for the cadastre and the other geoinformation sets of the INSPIRE directive [38].

With regard to the toponymy treatment, some works point out progress in applying new technologies. Among others, Conedera et al. prove that place names can also help find out past land uses, considering different, phonetically alike dialect options using a GIS [39]; Pijet-Migón and Migón examined the link between geoheritage and cultural heritage [40]. Furthermore, Blaschke et al. investigated how linguistic and cultural settings could affect the perception of place using a multi-language approach and a GIS [41]. Finally, Serikova and Baishukurova exposed an example of how Google Maps, Apple Maps, and Yandex can contribute to improving the exploration and identification of geographical names in Kazakhstan [42].

Moreover, toponym-comparison operations have been extensively analyzed for their use in search engines. In the case of the Canary Islands (Spain), the implementation of these solutions in GIS environments has been further developed [43]. Additionally, a general idea of the complete panorama of toponymy in Spain can be found in Gordón-Peral [44], as well as in the periodical conferences of the Commission on Geographical Names (*Comisión Especializada de Nombres Geográficos*). These works give brief indications of the sources used in inventory and research work, but do not go into procedures in any depth.

Finally, it is relevant to mention the management of toponymic gazetteer banks, both historical [45] and gazetteers of place names, in the frame of the European Union. Interoperability between gazetteers has significant advantages but also brings challenges with issues such as "localization, non-univocity, classification, level of detail or other data" [46].

2.2. On Gazetteer Synchronization

Attending to the digital treatment of geographical names as text strings, new technologies overcome challenges, such as those proposed by Deng et al., where the Levenshtein

distance is employed to compute the name similarity to conduct a geographical data integration in China [36]. On the other hand, toponymy can also be treated with natural language processing (NLP) technology [37,38]. Yan et al. made a deep neural network to consider both local and global features to deal with geocoding [39]. Martins employed machine-learning techniques to detect duplicated records in digital gazetteers [40].

These advances have a direct application in the synchronization of toponymic databases. In the case of the IGN (Spain), the synchronization of the state database (NGBE) with the regional databases has been tackled. This issue has been solved with their collaboration, using semi-automatic methods. In the international sphere, an example of the integration of large-scale databases is the United States federal gazetteer. This is called the United States Board on Geographic Names, and it combines local and state sources in its composition (https://www.usgs.gov/us-board-on-geographic-names; accessed on 2 February 2023).

A success story for detecting redundancies due to crowd-source contributions is reported from Indonesia. Gelernter et al. [47] address the same issue, comparing a created fuzzy match algorithm using machine learning (SVM). The latter checks both approximate spelling and approximate geocoding in order to find duplicates between the crowd-sourced tags and the gazetteer. In the Swiss case, two methods for removing duplicates are compared: the first method is based on rule-based matching, while the second applies machine learning using Random Forests [48].

2.3. On Cadastral Information and Cadastral Place Names

Certainly, cadastre has been understood in many ways, and in each territory has had its evolution, giving rise to studies on historical cadastre [49]. On the other hand, administrations have taken the path of standardization in this area [50], partly due to the development of technologies that make it possible [51] to move towards e-governance. Attempts to harmonize administrations have been crystallized in the ISO standard 19152 Land Administration Domain Model (LADM) [52], which in its work reflects the advances in data integration in Colombia. In the same way, synergies arise between different lines of technological development so that cadastral information is aided in the 3D concept supported by BIM [53]. In the case of Spain, the reform of the cadastral legislation by Law 13/2015, of 24 June, on the Reform of the Mortgage Law [54] has been a challenge for the administration and an opportunity for technical and conceptual improvement [55]. In addition to the works that study the cadastre, some investigations apply it as a tool for the study of other fields [56], such as urban planning [57] or historical studies [58].

Focusing specifically on the place names contained in the cadastral documentation, understood as geoinformation, which are present in numerous works is a vehicle for space interpretation. For example, one of the first cadasters in Spain dates from the reign of Fernando VI, the year 1749, called "Cadastre of the Marquis de la Ensenada", which was carried out in several places of the Kingdom of Castile, but not in the Basque Country because it was exempt from taxes [59]. On this dataset of great historical value, we can find lines of research under the linguistic prism [60], physical geography [61], interpretation as botany [62], and different applications through GIS related to the territory [63,64]. Specifically, cadastral toponymy can be helpful in the study of urban dynamics [65]. In another order, Pearn analyzes the relationship between the cadastral cartographic execution itself and place names [66]. Furthermore, there are studies dealing with the analysis of the authenticity of data recorded in cadastral databases [67].

Regarding toponymy gathered from cadastral information, academic documentation of the integration process is not frequent, but is an implicit part of the genesis of various datasets. Next, some examples of Spanish regions are presented. In the case of Aragon, according to the specifications of its gazetteer 2019 [68], the cadastre database has served as a source for the creation of the Geographic Gazetteer of Aragon, providing a large number of toponyms. In the creation of other regional databases, the cadastre is not cited. Yet the parcel structure of the cadastre is used as an element in the genesis of the place names. Specifically, Galicia has a very large volume of toponyms, partly explained by the

smallholding in land distribution [69]. This same element also appears in the case of the Balearic Islands. There, despite their systematic work based on field collection beyond the integration of cartographic sources, it is emphasized that the usual unit of nomination is the estate [70].

The toponymy of the cadastre has been taken into account in the preparation of local studies of toponymy, as in the case of San Martín de Unx, Navarra [71]. Even so, it has not been approached systematically throughout the territory, which is the aim of our study.

2.4. On Multilingualism

The geographic scope in which the study is contextualized includes a coexistence of languages, among which official bilingualism is found. Some authors prefer the term "languages in contact" to encompass all possible situations [72]. The current sociolinguistic context witnesses the presence of autonomous territoriality in which, in addition to Spanish and Basque, other languages of very different cultural origins coexist [73]. The political and identity issues related to multilingualism in place names have been studied by Jordan et al. [74] and Mácha et al. [7]. Moreover, the study of European policies and legislation is discussed by Ruiz Vieytez [75].

A key aspect for understanding the institutional management of bilingualism is the distance between coexisting languages [76]. In the case of Basque in its relation to Spanish, it does not enable semantic transparency between the languages. In this sense, we should remark on the emerging mainstream related to the study of the linguistic landscape [77], in environments of linguistic coexistence, especially in asymmetric situations [78], which deserves special attention. It even allows highlighting such asymmetry in urban space [79]. The contribution of linguistic varieties to forming an individual or group identity is explored in Rugkhapan [80], who makes special mention of the distance between Cartesian cartography and popular perception, official language, and everyday parlance.

Although linguistic standardization is a cutting-edge topic, the application of standardization in toponymy requires bearing in mind aspects related to its testimonial character [81]. In this way, "research is valued as a previous and essential step in toponymic standardization" [82], even more so in the case of Euskera, where the process of linguistic standardization continues to the present day.

3. Study Area

The geographical scope of this study is the province of Gipuzkoa (Figure 1), the smallest of the three territories of the CAPV (or CAE) in extension, approximately 1980 km^2 (https://ssweb.seap.minhap.es/REL/; accessed on 20 November 2022). This province has a population of 716,616 inhabitants (inh), which implies a population density of about 362 inh/km^2, slightly higher than that of its autonomous community, 302 inh/km^2 in 2022 (https://www.eustat.eus; accessed on 20 November 2022). This region's density is more than three times that of Spain, 94 inh/km^2 (https://www.ine.es; accessed on 20 November 2022), representing a high relative rate and leading to high pressure on the territory. In addition, Gipuzkoa has an irregular population distribution, depending on factors such as physical geography or industrial development. Specifically, the coastal regions account for three out of every four inhabitants [83]. Similarly, they are on the border of continental Europe, which is also a key factor [84].

Its regional organization is arranged according to the hydrographic basins or river valleys, most of which flow into the Cantabrian Sea, adopting different aspects of joint management of municipal services and collaborating in the structure of the territory [85]. Even though the secondary sector has a great weight, with the rise of the paper, textile, and, in part, iron and steel sectors, the area used for primary activities constitutes more than 60% of the total territory [86]. This last point is essential to understanding and justifying the study of the cultural heritage that toponymy represents, in any of its approaches, especially given its capacity to conserve the landscape information of the past [87].

Figure 1. The geographical extent of the research (province of Gipuzkoa, Spain). Source: Own elaboration from www.ign.es (accessed on 22 November 2022). Frame coordinates in km.

For the linguistic aspect, the Spanish Constitution [88] establishes a situation of co-officiality of Spanish and Basque throughout the autonomous community of the Basque Country. Furthermore, the Basque Government studies sociolinguistic dynamics through statistical operations such as the Sociolinguistic Survey or the Sociolinguistic Map. They show that 52.6% of the population knows Basque [89]. Despite this being the current situation, it should not be forgotten that the toponymy of any place is the result of multiple functional languages that have occurred throughout history, maintaining fossilized linguistic elements characteristic of past times [90].

4. Materials and Methods

In this section, the data used, the software utilized, the description of the process, and the validation strategy are explained.

4.1. Toponymic Geodata

4.1.1. Provincial Council. (Gipuzkoa)

The authority responsible for the management of geographic information in the study area is the Gipuzkoa Provincial Council, *Diputación Foral de Gipuzkoa* (DFG). Nevertheless, it has not developed a vocational toponymic database; it makes use of toponyms so that they can represent toponymic inventories. Its best-known geoinformation source, its 1:5000 cartography, shares a toponymic survey with the regional database, so a high correlation is expected. However, in this work, the source of information we want to give prominence to is the cartography of the Provincial Council's foral cadastre. It should be taken into account that the Cadastre of the Foral Deputations of the Basque Country, as well as that of the Foral (Autonomous) Community of Navarre, is outside the Spanish Common Regime territory [59], having had a separate trajectory historically with different data models. In addition to the characteristics of cadastral information, these data offer the particularity that they have not yet been systematically used to compile the official gazetteers. Hence, the contribution is original, expecting a lower correlation than among the rest of the spatial datasets. The cadastral information of the DFG is available online as described below: the parcel is divided into rustic and urban; on the one hand, there is a graphic description of the parcel (https://www.gipuzkoa.eus/es/web/ogasuna/catastro/informacion-general; accessed on 29 November 2022); on the other hand, there is an alphanumeric description (https://www.gipuzkoairekia.eus; accessed on 29 November 2022).

4.1.2. Autonomous Community (Basque Country)

The toponymic corpus of the Basque Government acquires its current status through the Decree 179/2019 of 19 November, "on the standardization of the institutional and administrative use of the official languages in the local institutions of Euskadi" [91] under the protection of the European Charter for Regional or Minority Languages [92] (instrument of ratification by Spain, "Spanish Official Bulletin"—*Boletín Oficial del Estado*—of 15 September 2001 [93]). This is because, as indicated in its preamble, under the UNESCO Atlas of the World's Languages in Danger, Basque is a minority and vulnerable language in its own territory and is in an asymmetrical situation concerning the other official language, namely, Spanish. The fifth chapter of the aforementioned regulation deals with municipal toponymy. It pays particular attention and respect to the distribution of competencies and responsibilities of different public authorities, including those of the Academy of the Basque Language, *Euskaltzaindia*. In addition, the NGO-CAE is created as a public register, attached to the department with competencies in matters of linguistic normalization of the Basque Government, in which the official place names of the Basque Country will be registered.

The precursor of this database is the agglutination of different works of collection and processing of place names, which have been deployed since the 1990s [94]. These have been carried out on the initiative of the autonomous administration in its territorial scope and, additionally, different local projects that have been complementing, improving, and updating the original database. The autonomous government itself has used different measures to boost this local work. Together with regional and local contributions, this corpus has a linguistic vocation that prevails over the geographic component. Thus, the intended use of this dataset within the institution's cartography grants it the nature of geoinformation in all its functionality. This research adopts the compact version of the NGO-CAE (https://www.opendata.euskadi.eus/inicio/; accessed on 29 November 2022), and materialized in a text file, and presenting point geometry. This characteristic simplifies the vector overlay geoprocessing and reduces processing time. In case of the need to retrieve the geometry of the features, it is possible to link the "identity" field of this database with the identifier of the toponym collected in the Harmonized Topographic Base, *Base Topográfica Armonizada*—BTA.

4.1.3. National Level (Spain)

At the national level, the National Cartographic System establishes the National Geographic Gazetteer as part of the National Reference Geographic Equipment, which is made up of the harmonization of the NGBE and the Geographic Gazetteers of each of the Autonomous Communities [95]. Therefore, the NGBE has been completed by the IGN to meet the requirements of the INSPIRE directive and the LISIGE. Although the starting point of the database is the toponymic content of the *Mapa Topográfico Nacional*, 1:25,000 (MTN25) and the rest of the IGN cartographic series, the maintenance and improvement activity over the last decade has been aimed at coordination between gazetteers. After a process called "autocorrection", the correction of NGBE toponyms by comparison with the NGO-CAE was addressed through a process with an extensive manual component [17]. Thereby, the strong correlation or convergence between these two databases is recognized in the case of the CAE. On the other hand, this database, managed from a regional perspective, oriented to cover the needs of cartography at a scale of 1:25,000, has a much lower density, so its contribution is limited.

4.2. Software Utilized

Data management and pre-processing, as well as the comparison of their contents, including different linguistic aspects, was performed using ETL tools, specifically *FME© Desktop 2021.2* (https://www.safe.com/fme/fme-desktop/; accessed on 30 November 2022). Alternatively, the analysis of the results and the methodology validation were conducted using a GIS tool, *QGIS 3.24.0 Tisler*; www.qgis.org (accessed on 30 November).

4.3. Processing Overview

The process followed can be summarized graphically using the diagram shown in Figure 2.

4.3.1. Input Data

First of all, in the NGO-CAE a series of operations was performed on the data to keep only the objects of interest, eliminating repeated elements that were due to entries with a larger set of attributes that differentiate each instance. This issue is conditioned by the original structure of the Vice-Ministry of Linguistic Policy's own database. It was also necessary to eliminate the coding of communication routes, which, as a literal system of geographic location, in contrast with toponymy, has a normative and not a cultural origin; therefore, this information was discarded. The input of these data in the ETL processing was through an alphanumeric file (in .CSV format). The comparison field we used contained the specific part of the name to avoid as much as possible the coincidences that would occur if we included the generic part. In this sense, the gathering capacity of the Basque language, which favors the agglutination of lexicalized generics, caused the appearance of coincidences of lexemes specific to toponymic or geographic terminology.

In the cadastre-DFG database, numeric codes are also inserted together with the toponymy (for example, the portal number of specific parcels). Still, their elimination was incidental since they did not affect the processing. The coding of the parcels in polygons, together with the municipality code, facilitated the union of the geometric element representing the parcel and its alphanumeric attributes, including the place name.

For the comparison of the elements of each origin, for computational efficiency reasons, the centroids of the cadastral parcels (CP) were superimposed on areas of influence around the NGO-CAE place names. For areas where the parcel extension was large, it was decided to generate a continuous surface of the Voronoi diagram, as has been used in other recent cartographic works [96–98]; these are usually forestry and pasture areas. In most cases, the municipality owns the land, either as public domain or as patrimonial land.

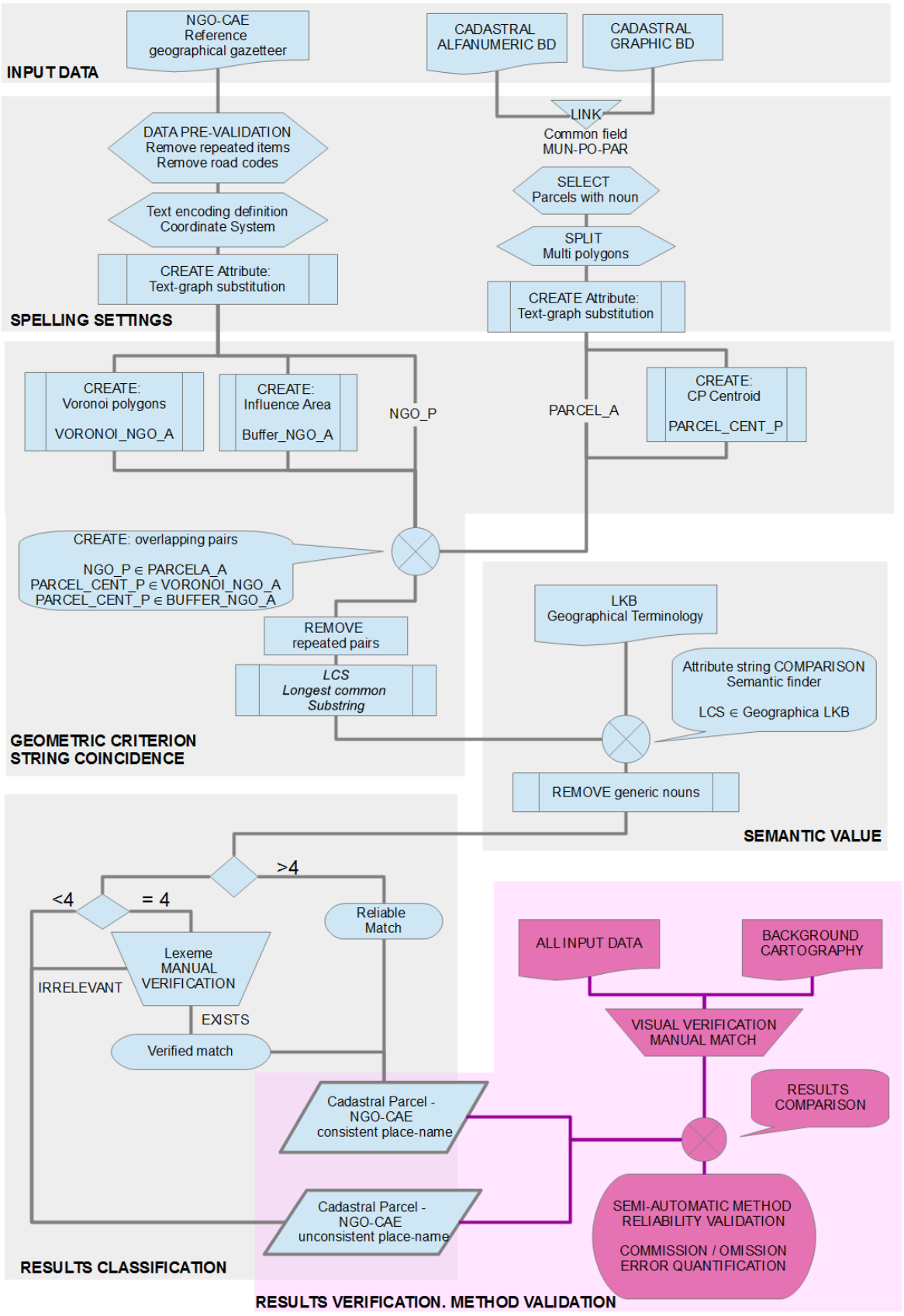

Figure 2. Flowchart with the process undertaken.

4.3.2. Spelling Settings

In this work, we intended to address the complexity arising from the different states of linguistic standardization of the databases in advance. Therefore, the field containing the place name was not compared, but a previous transformation was applied to avoid this issue. In the spelling field, the procedure was designed to transform place names to a comparable state, paying particular attention to the writing of sibilants, which have seen different forms of transliteration over time. Consequently, for the purposes of this study, they are represented graphically by a single sign. In the same way, it was necessary to obviate capital letters since, both in cartographic labeling and in literal database management, different streams are followed. Among these streams, we can mention the following: IGN standards of capital letters as a title but with variations depending on the lexical category of each word; French standardization; or the current trend of the Basque Government in which the specific name is in upper case and the generic in lower case; also in the cadastre, only capital letters are used, with different applications of the rules of accentuation. In addition, in order to avoid situations in which the Spanish spelling was used in names with Basque roots, the letters "C" and "K" were replaced by a single sign. To the extent that these differences were previously mitigated, the comparison was more efficient. These issues have been tackled using neural networks in Alexis et al. [7], both in onomastics and toponymy, as well as in GIS-based applications, such as the one described by Dalvi et al. [99].

4.3.3. Comparison Premises

Therefore, the comparison had multiple components. Firstly, the geometrical aspect; the elements compared had to have geographical proximity; next, the string coincidence or character proximity, which means text similarity; finally, the semantic component, which improved results regarding the meaning (generic names).

Geometric Criterion

On the one hand, the parcels had geometry and position in space since they were surface elements. Still, the parcel size has an influence on the evaluation of the distance with the NGO-CAE elements, as in the case of parcels with more extensive land use, such as agricultural or forestry activity. On the other hand, in the case of toponymic databases, it must be taken into consideration that some geographical elements do not necessarily have an unequivocal position on the cartography, or have poorly defined physical boundaries, as is the case of the sites. In addition, the source for the formation of the gazetteers is the cartography itself. Therefore, the toponyms' locations may have been altered by the needs of the cartographic compilation process or its subsequent digitization. For these reasons, we decided to define a large enough buffer, and an overlay process was implemented. Hence, the cases of nominated parcel centroids (the parcels with an empty "name" field are left aside) contained in the buffer of each NGO-CAE element were passed to the alphanumeric comparison.

For the purpose of defining the size of the buffer, two parameters were accounted for: the average size of the parcel and the length of the label according to cartographic technical specifications [100], obtaining a value of 120 m. Based on this initial assessment, the need to increase the size is empirically verified to also cover the overlap between large-surface features such as sites (*parajes*), which in the NGO-CAE are represented by their centroid. In order to decide the final size (200 m), it was ascertained that the processing time remained stable on the sample space used for the evaluation of the method. On the other hand, tests with a buffer larger than this value did not yield appreciable improvements in the number of matches.

String Coincidence

The comparison of the text string of the field containing the name in each data source was undertaken. In this stage, we first chose the field in each database in which the generic component of the place name did not appear. Subsequently, the necessary modifi-

cations were applied to solve the issues of spelling and normalization of the algorithms related to the proximity between character strings. Under these circumstances, it was found that the most efficient way to find common lexemes in the comparison was that of "longest_common_substring" (LCS) [101]. Appendix A contains the code of the LCS subroutine implemented in Python language (scripting compatibility 3.8+).

This algorithm operates as follows: the content of the strings of both place names (NGO-CAE and cadastre) of all the pairs in which geometric proximity was verified was compared letter by letter. The composition of the common string accumulated successive identical characters. It stopped either at the end of one of the comparison strings or when a different character was found between them. In the latter case, it continued searching, composing a new substring. Later, it selected the one with the largest length from the identical substrings. The result was independent of the position of the substring within the place name, whether it was the beginning of the string, the end, or the intermediate part. The result obtained from the comparison was the longest common string of characters, and its length, as an integer. For each cadastral toponym, there was a set of NGO-CAE candidates obtained by geometric overlapping, which could be ranked according to the results of the LCS application.

For the sake of clarity, we present two examples. The first case dealt with the parcel "Ibiri Aldeko Sakona" (site), where matches were found. The longest match was Ibiri (the name of a *Caserío*). This led to a successful coincidence. In another example, the parcel "Estanga Haundi" had several matches with the common string "Estanga", such as Estangagoena, Estanga Bekoa, Estangabarrena, Estangaundia. This constituted another satisfactory example; still, it also matched another string ("Urkolaundia") through a descriptor particle ("aundia" or "handia" means large). Hence, this coincidence had no semantic justification for being rated as a valid match.

We preferred this algorithm to others because of the variability in place name formation. For example, in the case of the proximity of the entire string using an index, such as "fuzzy string matching" [102], agglomerated lexemes can come into play as part of the name that masks the result obtained for the proximity or similarity index. The output was a list of pairs of elements (parcel-gazetteer) that could be clustered according to the number of characters they have in common. It only remained to verify that the pairs repeated and to analyze the results.

Semantic Value

The complexity derived from the mere definition or composition of place names also needed to be faced. Thus, they could be found in an extended form if they were made up of their generic part in addition to the specific part, or the specific part could be recorded exclusively. Moreover, in the semantic field, the same root lexeme could be the origin of a family of names, but at the same time, they could allude to different entities. Finally, it was possible to find translations in both the generic and the specific part of the place name.

In order to be able to interpret these questions and thus verify the proximity between texts both phonetically and semantically, it was necessary to compare the coincidences with a geographical glossary. The latter refers to an inventory of frequent lexemes in geographical elements or their description. This would be a "Lexical Knowledge Base" (LKB) oriented exclusively to geographical concepts [103]. This LKB is based on that published in Azcárate Luxán et al. [104], a document currently being revised and updated. As the content of the glossary increases, the subject matter of each name can be semi-automatically profiled.

4.3.4. Result Classifications

From the application of this procedure, a series of pairings between the names of the two source databases was obtained, with the longest common substring and this substring length. Results verification was performed by manual review of these links in a sample space and is explained in Section 4.4. Given the heuristic nature of the problem, this sample

allowed inferring the threshold for the minimum length at which a match is a semantic unit of the specific part of the name. We estimated this threshold as four characters.

4.4. Results Verification (Method Validation)

The validation of the method was performed by checking the results against a manual run. First, data from four municipalities with different characteristics were taken to obtain a representative sample space of the cadastral parcel names. The municipalities selected are Abaltzisketa, Oñati, Ordizia, and Mutriku (Figure 3). Based on population density extracted from the 2022 Eustat data (www.eustat.eus; accessed on 2 December 2022), Abaltzisketa was chosen as a sample of territory with low density (26.43 inh/km^2; Oñati, with a population density of 106.46 inh/km^2 and an economic activity polarized between the primary and secondary sectors; and Ordizia, with extreme demographic pressure on the territory (density of 1762.32 inh/km^2); in addition, since the selected municipalities are inland, a coastal municipality was added, such as Mutriku, with a density of 193.12 inh/km^2.

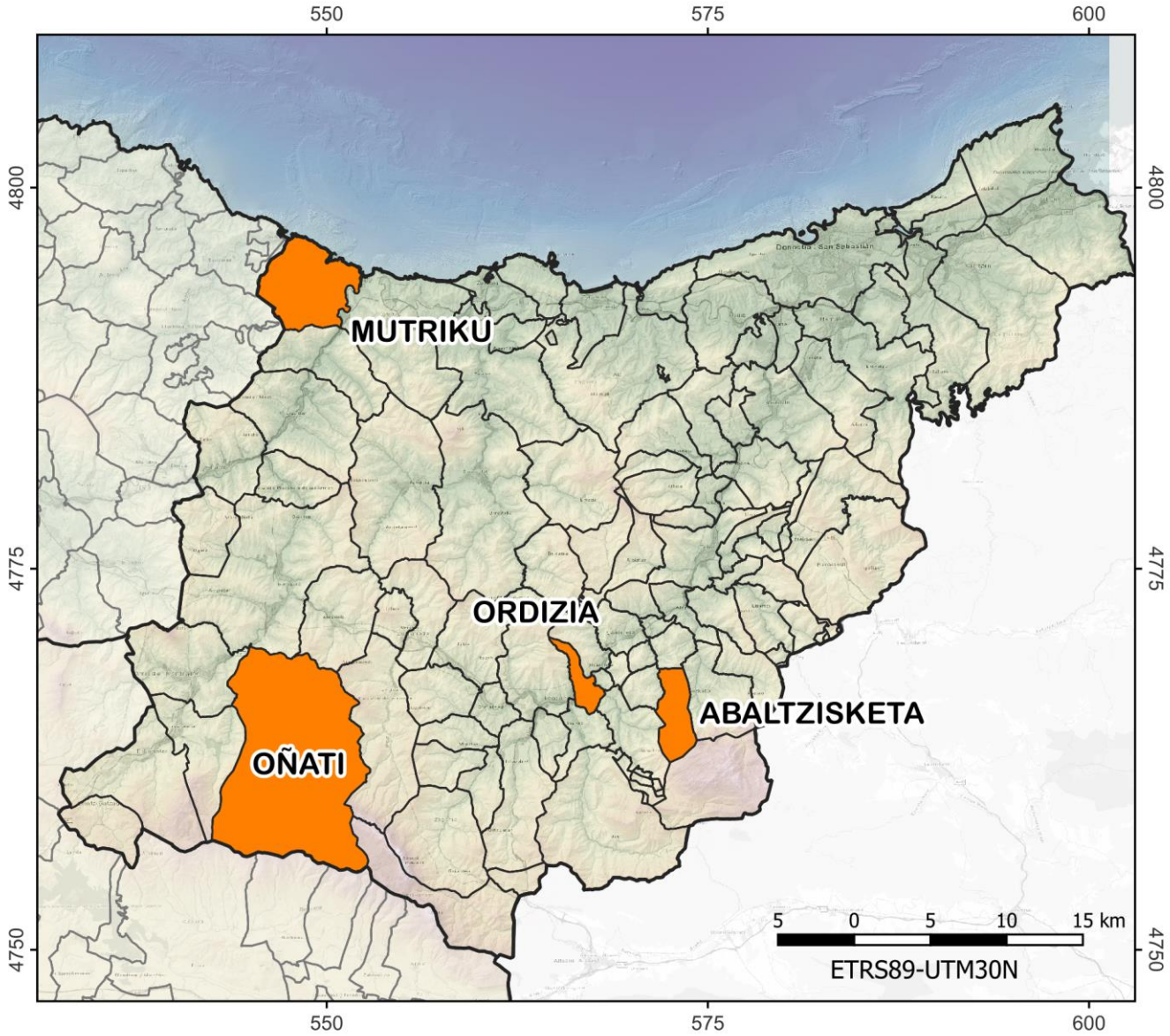

Figure 3. Location map of the sample municipalities. Frame coordinates in km.

The pairs obtained and hierarchized using the methodology described above were validated, checking the degree of semantic proximity of the combinations obtained. After

analyzing the LCS strings obtained and matching them with the glossary of geographical terms, a list of lexemes and text strings was created to filter the total number of pairs determined so that the final result was close to the manual review.

5. Results

This section includes the results generated in the proposed process and their analysis. The methodology implemented and tested in this study has been designed to verify the coherence and extract complementary information regarding place names from the geographic information of the cadastre of the province of Gipuzkoa. The reference database used is the NGO-CAE. Table 1 shows a quantitative evaluation of the two sources of information: the municipalities selected as a sample and the province as a whole. Out of the 115,282 parcels depicted in the cadastral dataset, 45,326 contain an entry in the name field, representing 39.3% in relative terms. In the case of the NGO-CAE, the number of elements is 25,108 after eliminating repetitions and road codes.

Table 1. Number of elements stored in the databases under analysis.

Administrative Unit	Cadastral Parcels (CP)	CP with a Name	% CP with a Name	NGO-CAE
Abaltzisketa (001)	1050	508	48.4	160
Mutriku (056)	1415	556	39.3	396
Ordizia (079)	426	111	26.1	70
Oñati (059)	7235	2661	36.8	1331
GIPUZKOA	115,282	45,326	39.3	25,108

From the process execution to search matches between toponyms in the two databases, a list of pairs of elements is returned, classified by the length of the concordance chain (Table 2). The number of matches for two-character substrings is very high, 450,736 matches, and decreases with an increasing number of characters, as is the case for the 26,626 matches for five characters or more. The longest match found is 20 characters, specifically from the parcel coded 074-04-015, called "Zezeagako Errekaldea." Each parcel can have more than one match, even with the same NGO-CAE toponym. This is not an obstacle for a single coincidence considered reliable or validated to be sufficient in the verification of the redundancy of the databases. Furthermore, in the same way, in the case of not finding reliable or verified matches, the name of that parcel will be considered not registered in the NGO-CAE. Therefore, it will be considered as a potential contribution to the toponymic corpus resulting from the integration of sources.

Table 2. Number of data pairs found using the "longest_common_substring" (LCS) operator, sorted by string length.

Administrative Unit	6C+ Long LCS	5C Long LCS	4C Long LCS	3C Long LCS	2C Long LCS
Abaltzisketa (001)	221	113	664	1696	5359
Mutriku (056)	111	338	314	1278	6361
Ordizia (079)	108	63	92	291	2254
Oñati (059)	1122	512	2103	6250	29,510
GIPUZKOA	18,171	8455	32,879	99,277	450,736

The next step is to assess quantitatively how many of these combinations have a real significance and how many are mere coincidences of character sequences, due more to chance than to their non-existing semantic link. As the lexemes are collected in groups, this review is relatively agile if we organize the database by the "common_substring" field, since the lexemes are collected in groups. This check also detects those substrings that do not constitute lexemes, such as declension marks, adjectives, or generics that accompany the proper name. The case of adjectives is very common in the names of buildings. For

example, the building "Azkarraga Haundia baserria", or "Large *Caserío* Azkarraga" with its adjective "Large" probably differentiates it from other buildings with which it shares the lineage or root "Azkarraga". Regarding the hydrographical generics, we found that in both punctual and linear, the tendency is to agglutinate the generic, so in the process, coincidences with the word "erreka"—creek—and "iturri"—fountain—are introduced as successful cases. This issue underlies the need to incorporate the semantic criterion by contrasting it with a geographic LKB.

From this review, we estimate the number of characters necessary to obtain a meaningful match. As long as the length of the matched string is greater than four characters, complete lexemes, endowed with semantic content, are found. This is the reason why the data relating to matches equal to or greater than five characters are reflected, grouped in a single set. When the string length is four, the volume of matches that do not correspond to semantic units grows so it is only possible to validate that these pairs are meaningful with a manual process. Below this substring length, the use of LKB does not provide sufficient reliability, and manual verification is so costly that it is not contemplated in this study. As shown in Table 3, the results for the group of strings of five characters or more (5C+), the verified matches (Figure 4), and those obtained after comparing the total with the LKB, present similar results (Figure 5). Still, a more conservative result has been reached for the automatic method to reduce errors of commission.

Table 3. Results obtained for strings of five or more characters; totals, with supervision and automated matching with LKB. Figures are broken down into number of matches and number of parcels.

Administrative Unit	5C+ Match	5C+ Match Parcels	Significant 5C+ Match Supervised	Significant 5C+ Match Parcels	5C+ Match and Semantic Filter (SF)	5C+ Match and SF Parcels
Abaltzisketa (001)	334	157	294	130	286	118
Mutriku (056)	449	256	373	210	343	214
Ordizia (079)	25	20	23	20	23	20
Oñati (059)	1689	1125	1424	1035	1424	972
GIPUZKOA	27,679	17,306			22,617	14,691

Table 4 displays the four character results, both supervised (Figure 6) and automatic (Figure 7). As can be seen, the depuration of the results with the LKB has not assured the automatic method's reliability. Furthermore, as inferred from the last column, some municipalities have commission errors, such as Abaltzisketa or Mutriku. For these reasons, a manual check of this dataset is necessary. Finally, regarding the efficient planning of the process, it should be emphasized that if from the total of 4C combinations found, we exclude those related to plots already combined in the previously described step, the amount of data to be processed decreases substantially.

Table 4. Results attained for four-character strings; totals, with supervision and automating the matching with the LKB. Figures are broken down into number of matches and number of parcels.

Administrative Unit	4C Match (without 5C+ Parcels)	4C Match Parcels	Significant 4C Match Supervised	Significant 4C Match Parcels	4C Match and Semantic Filter (SF)	4C Match and SF Parcels
Abaltzisketa (001)	664	172	64	23	47	38
Mutriku (056)	223	90	46	22	72	44
Ordizia (079)	29	10	7	1	14	4
Oñati (059)	150	50	28	10	40	19
GIPUZKOA	18,092	7128			7210	3488

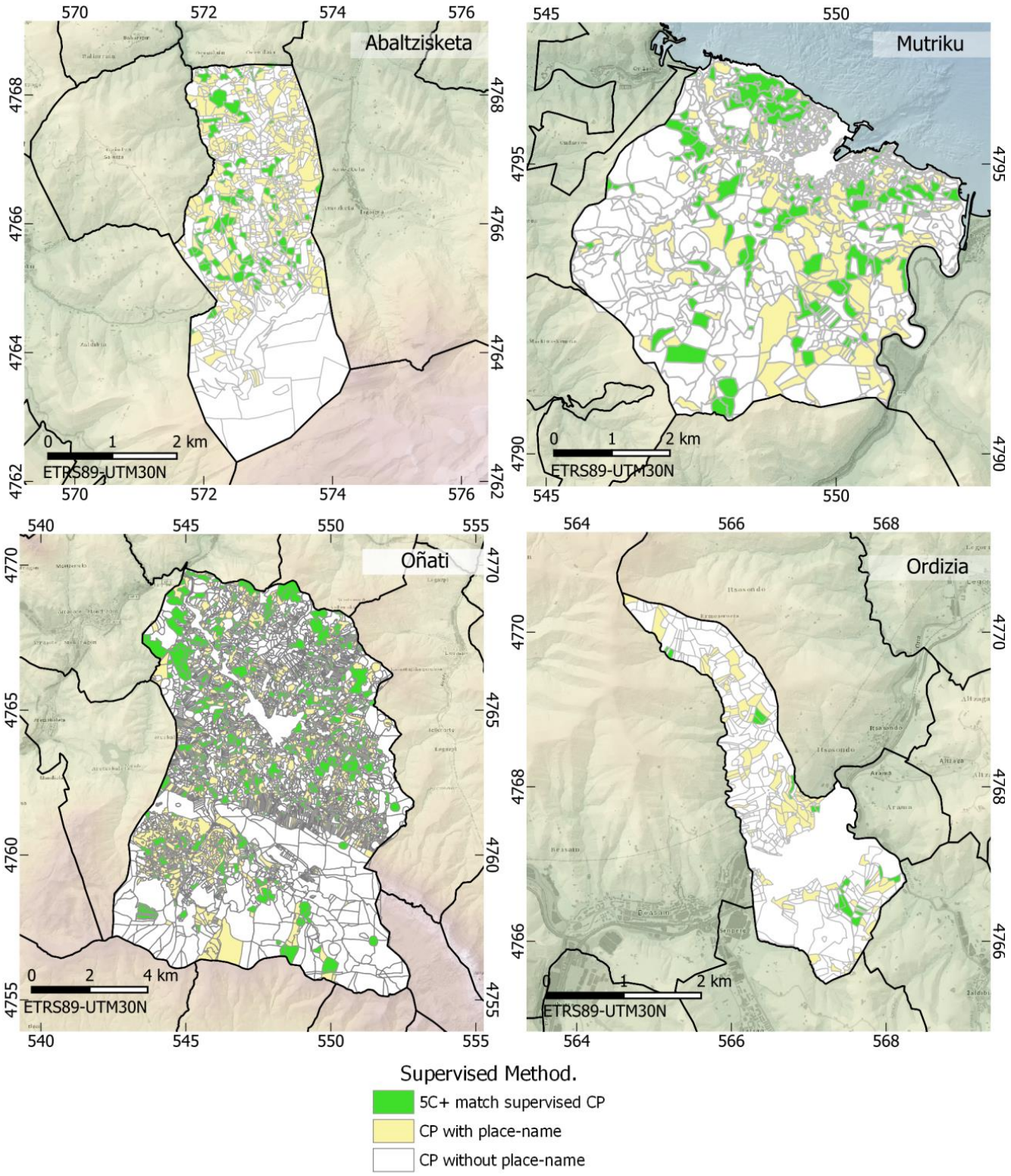

Figure 4. Cadastral parcels selected by supervision (5C+ matching). Frame coordinates in km.

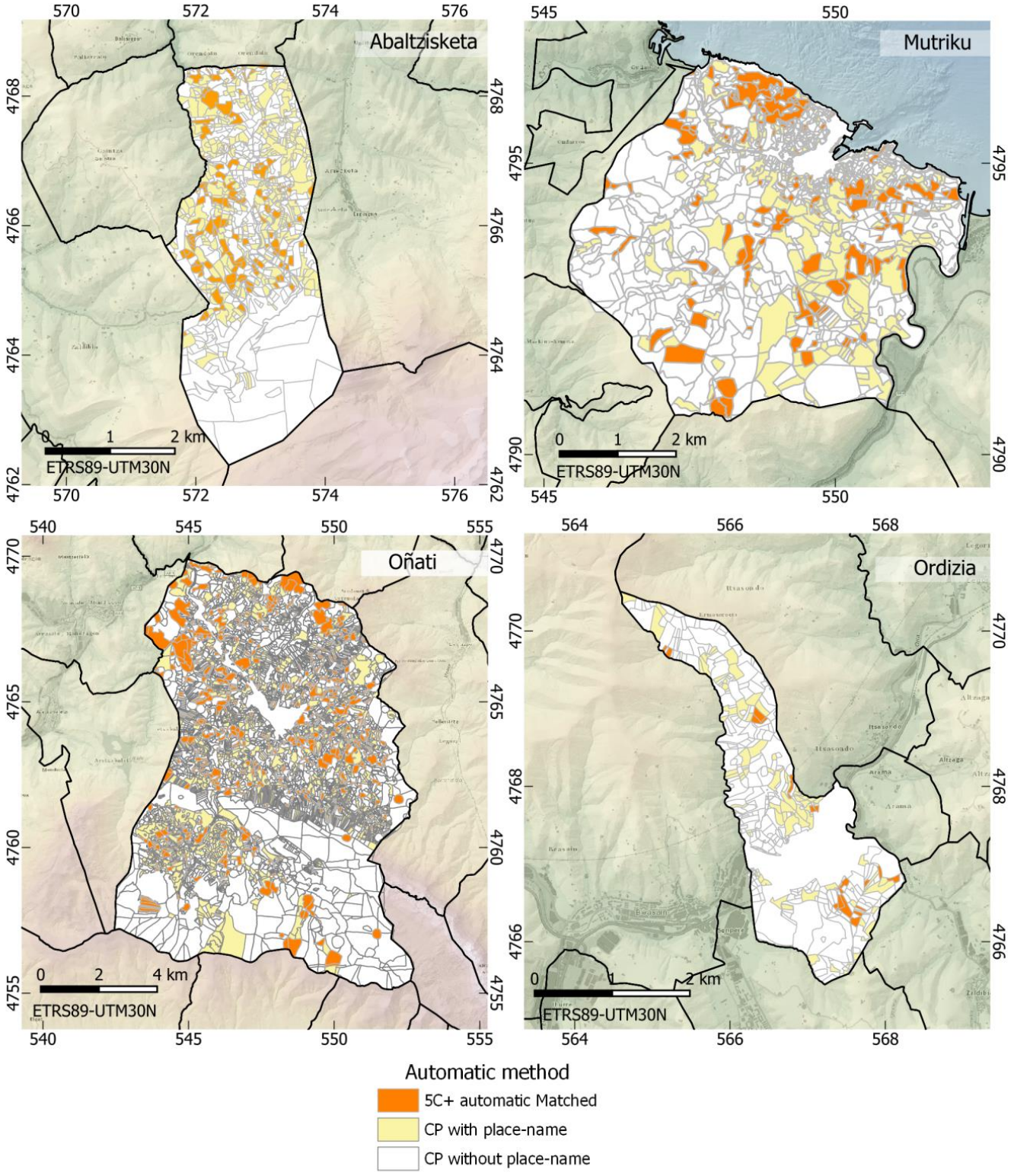

Figure 5. Cadastral parcels selected automatically (5C+ matching). Frame coordinates in km.

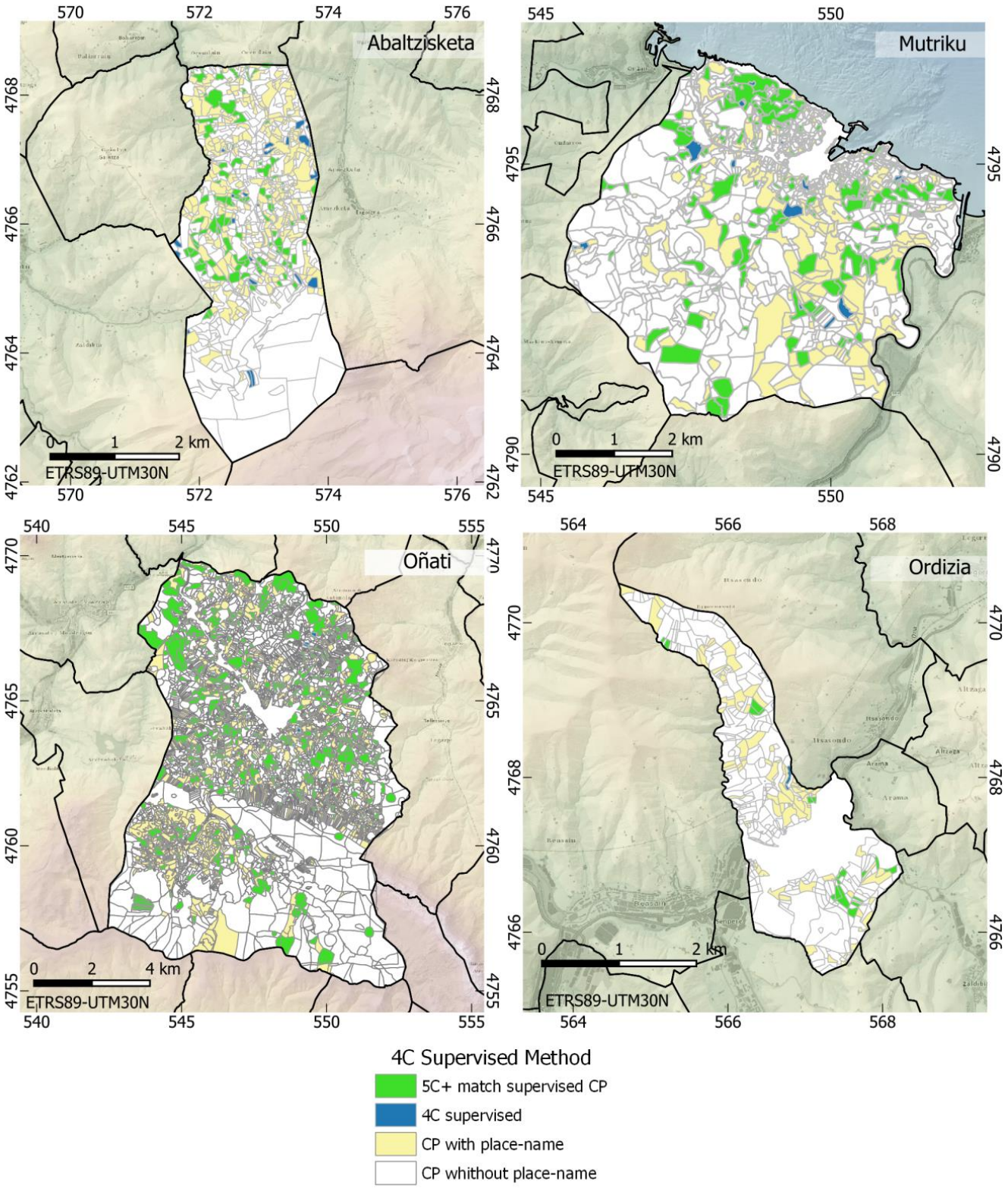

Figure 6. Cadastral parcels selected by supervision (4C and 5C+ matching). Frame coordinates in km.

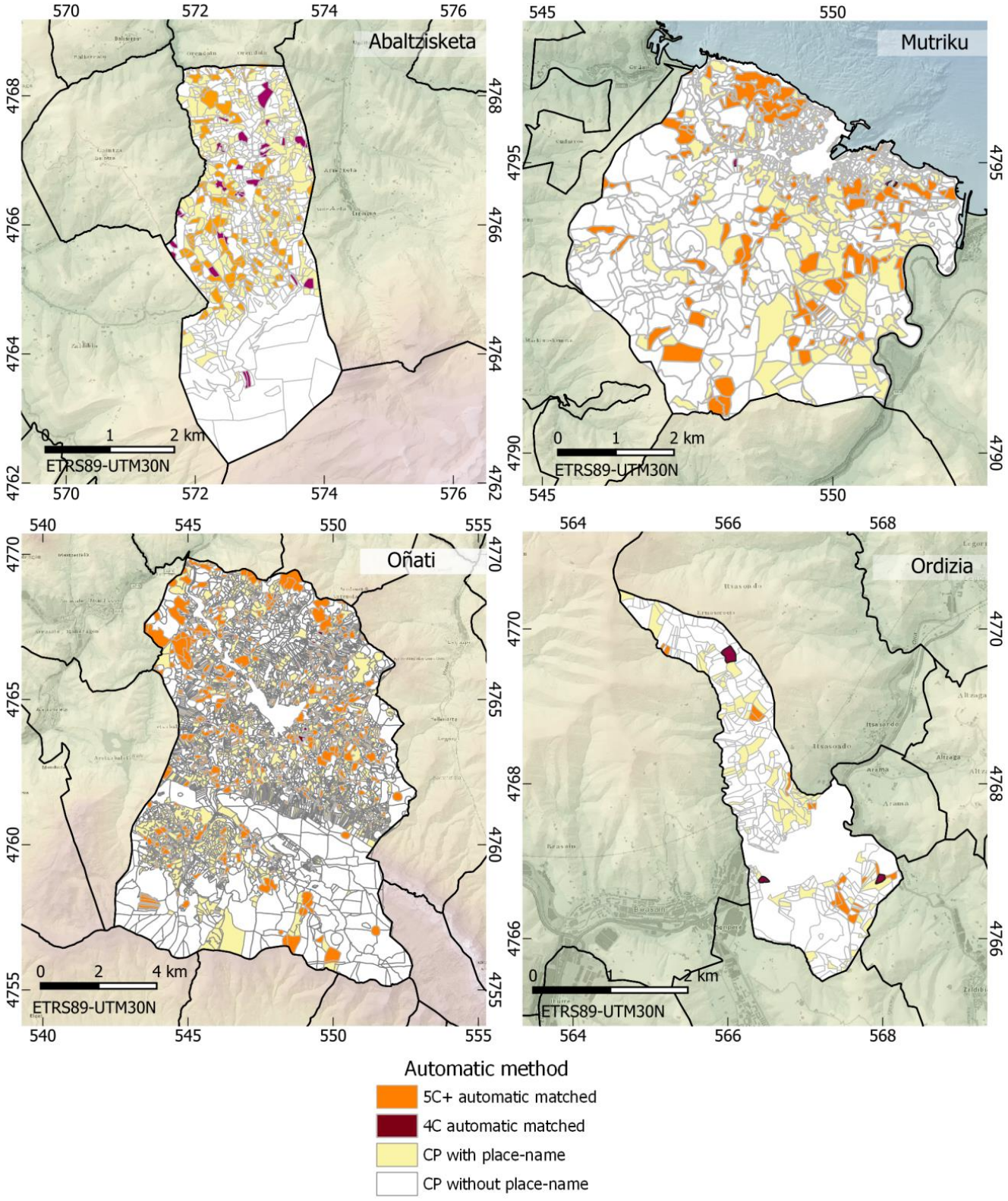

Figure 7. Cadastral parcels selected automatically (4C and 5C+ matching). Frame coordinates in km.

Finally, to derive a quantitative assessment of the quality of the process, Table 5 reports the results of quantifying the discrepancies between the automatic method and its manual verification (Figure 8). As a result, a rate for errors of omission and errors of commission is obtained for the sample. As explained above, semantic matching on the sample has been

used to demonstrate the reliability of the matches found automatically. In the case of error of omission, a rate of 5.7% was observed. This value is calculated on the nominated parcels of the sample space. Similarly, the commission error rate is 1.7%. This percentage can be modulated by applying the lexical filter more or less restrictively, but reducing the rate of errors of commission inevitably means raising the rate of errors of omission.

Table 5. Revision of errors of omission and commission in the sampled municipalities.

Administrative Unit	CP with a Name	Total Manual Significant Match CP	Total Auto + Semantic CP Selection	Omission Occurrence CP	Commission Occurrence CP
Abaltzisketa (001)	508	180	156	18	3
Mutriku (056)	556	346	258	25	23
Ordizia (079)	111	21	24	0	0
Oñati (059)	2661	1045	993	176	39
% Total (from Total CP with manual significant match)				13.8%	4.1%
% Total (from Total CP with a name)				5.7%	1.7%

To illustrate this issue with an example, we use the adjectives "zabal" (wide) and "sakon" (deep) in the sample of the municipality of Oñati. For the first of them (zabal), this coincidence is automatically detected 32 times. After manual supervision, 19 elements are identified as significant in concordance with the specific; on the contrary, 13 are not. Because of these results, it is concluded that the studied substring is not a discriminating element in the consistency of proper names and is not applied in the filtering. Conversely, 18 automatic matches are found with the adjective "sakon", and the manual review reveals that only two are inconsistent. In this case, deciding to include this lexeme as a discriminator depends on how conservative we want to be with commission errors.

To conclude, after checking the method's validity, Table 6 summarizes the results derived from applying the method to the universe of discourse (Figure 9). Finally, the occurrence rate detected by the two methods for the sample and using the automatic method for the whole province of Gipuzkoa is presented. Out of the 18,179 cadastral parcels (CP) nominated, 40% have some element that is partially or totally coincident with the NGO-CAE.

Table 6. Results obtained after the application of the automatic process in the total population (on the parcels of the entire municipality).

Administrative Unit	CP with a Name	Total Manual Significant Match CP	Total Auto + Semantic CP Selection	Auto + Semantic % from Significant Match CP	Auto + Semantic % from Significant with a Name
Abaltzisketa (001)	508	180	156	86.6%	30.7%
Mutriku (056)	556	346	258	74.5%	46.4%
Ordizia (079)	111	21	24	114.3%	21.6%
Oñati (059)	2661	1045	993	96.0%	37.3%
			Sample average	89.9%	37.7%
GIPUZKOA	45,326		18,179		40.1%

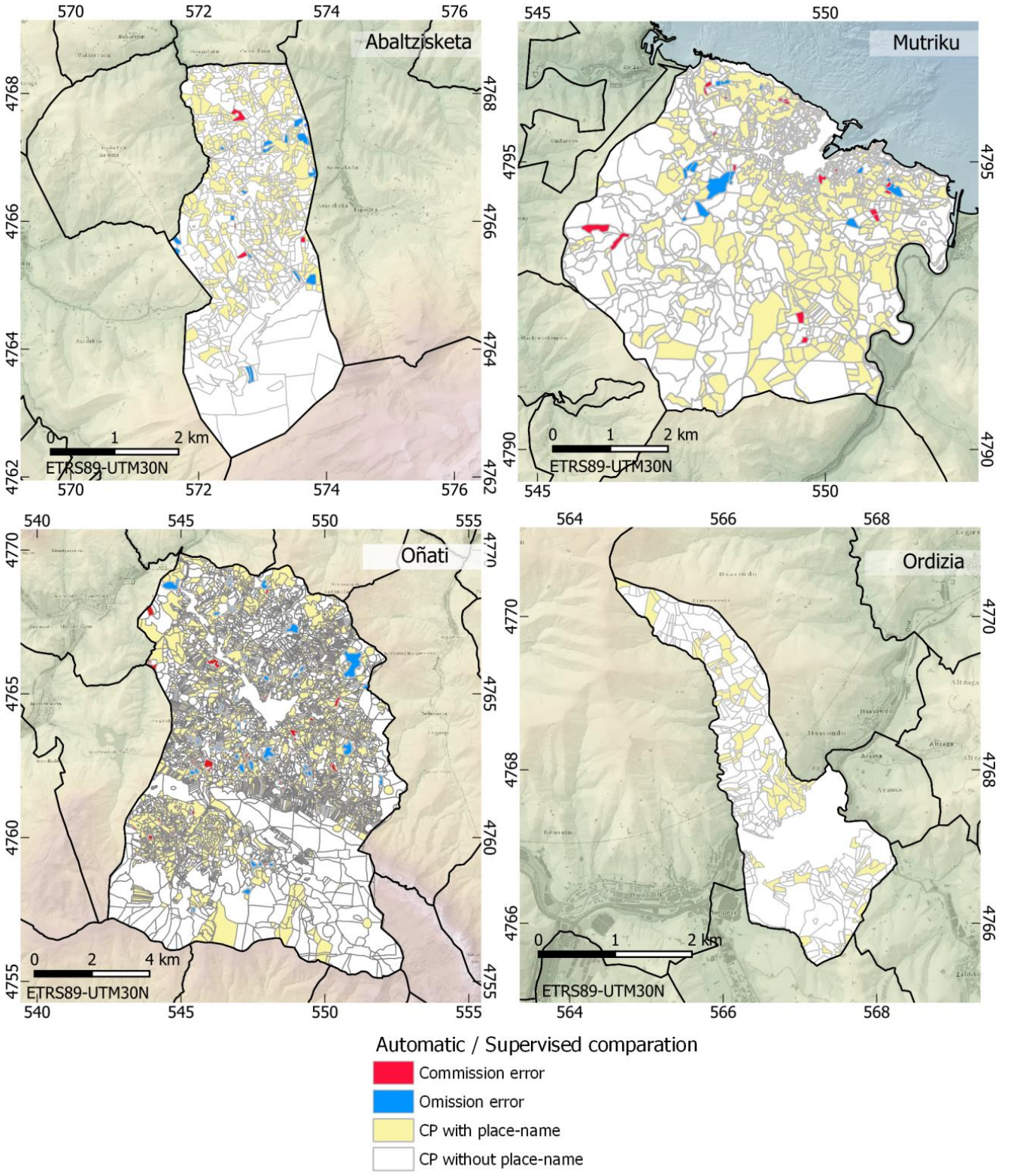

Figure 8. Omission and commission errors in the selected cadastral parcels. Frame coordinates in km.

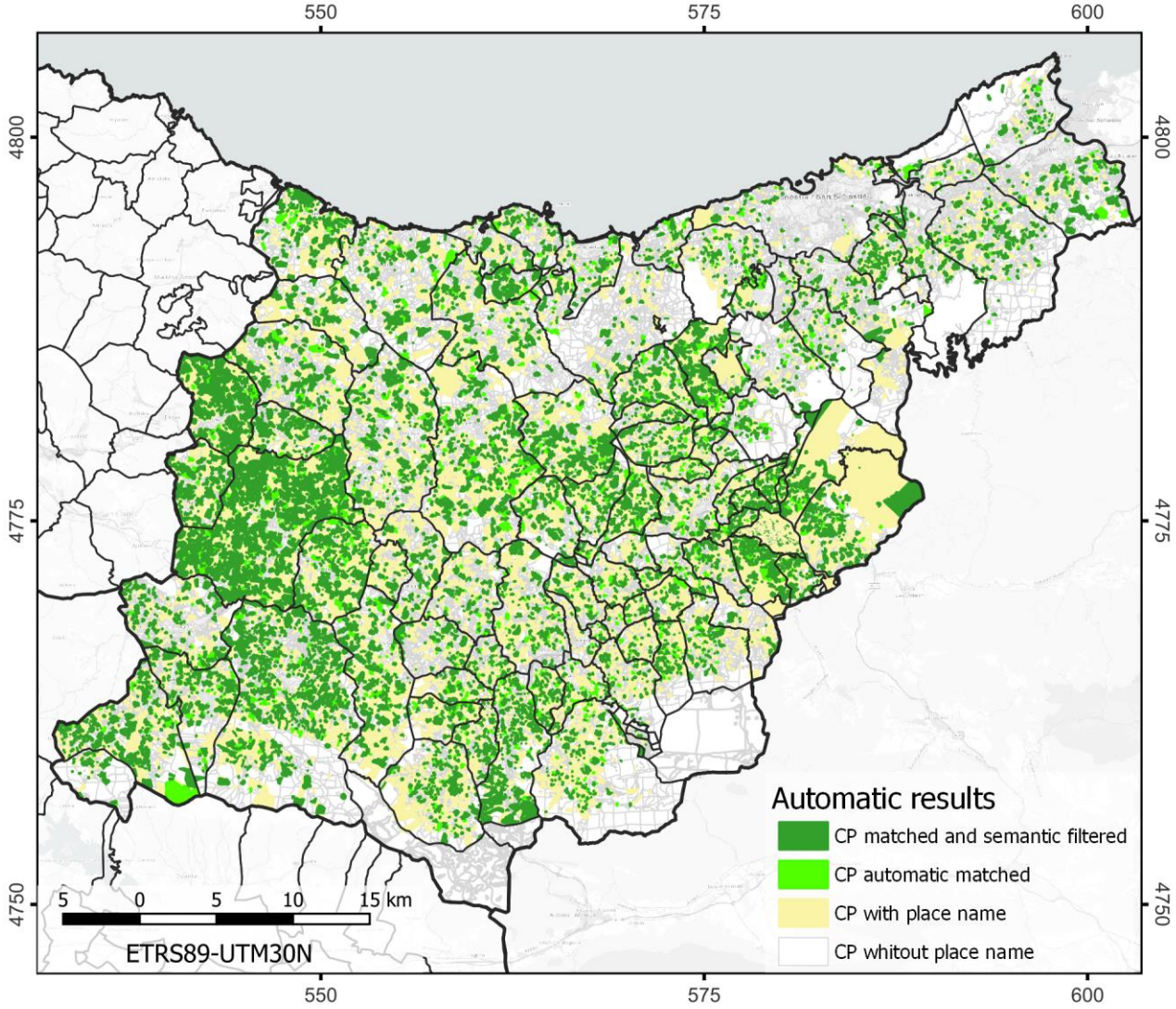

Figure 9. Results from the proposed method to the whole province of Gipuzkoa. Frame coordinates in km.

6. Discussion

In this work, we have designed a process using ETL and GIS tools to assess the redundancy between two groups of toponymic data originated from the geoinformation of different natures. The objective is their integration and joint use in subsequent spatial analysis works. In pursuing this redundancy, the criteria of location and textual coincidence have been applied, as well as the semantic criteria. The methodology has been applied, providing absolute and relative values that allow us to positively evaluate its usefulness and the use of the sources, as argued below.

Firstly, we have designed an automatic process, which has attained a concordance of 40% for coincidences in the texts of the five-character names. The result is susceptible to improvement with semi-automatic processes, such as including the results of four-character matches, which would reduce the 5% value of errors of omission. Redundancy can also be interpreted in a complementary way; it can be estimated that 60% of the toponymic corpus is susceptible to being integrated into the NGO-CAE. Apart from considerations about the feasibility and provenance of this task that we will discuss below, we can compare these results with the analogous work carried out by the Andalusian Regional Government [16].

While the contribution of the cadastral toponymy to the Andalusian Gazetteer (NGA) accounts for 18.7% of elements, with the considerations presented below, we can calculate

that the contribution to the NGO on the territory of Gipuzkoa could reach up to 46.9%. This estimate is derived by considering that the 18,179 concordances of the result (which represent 40%) have been established with 7850 elements of the NGO-CAE, which is 31.3% of the gazetteer. This is explained by the fact that there is also an internal redundancy or correlation between the names of the parcels. Therefore, following the same logic and applying this redundancy rate to the 60% of parcels with independent names, we arrive at the figure of 46.9%. In any case, this is only a maximum estimate, but a substantial improvement concerning the results of the Andalusian project cannot be ruled out. This is because, while the databases compared in Andalusia do have interdependence in their content [16], this is different in the Basque Country. In other words, while in the initial specifications of the cartography *Mapa Topográfico de Andalucía* (MTA), the cadastral cartographic information between the years 1985 and 1990 was established as a source or reference [16]; conversely, for the preparation of the BTA in the Basque Country the toponymy of the foral cadastre has not been used [105], which means a lower correlation. More specifically, a correlation exclusively based on the object and not on the process is expected.

On the other hand, it should not be forgotten that the most significant difference between these two works is that, while the improvement of the NGA is performed in the administrative field, the initiative of our study is academic. The requirements for the inclusion of the official geographical gazetteer of the Autonomous Community of the Basque Country are regulated in Decree 179/2019. Under this regulation, the initiative must come from a competent body, which is the local administration in the case of microtoponymy. In addition, this process requires the classification of candidate names based on the target data catalog. Thus, this task transcends the scope of this work and academic research. Nevertheless, this is not an impediment to considering this integrated dataset in developing works of analysis of the territory based on place names. In this case, the integration of sources can only further the process and increase the robustness of the tests.

Another project related to this study, framed in the National Geographic Gazetteer (NGN) preparation, is the correlation of the initial version of the NGBE with the toponymic databases and information of the autonomous communities developed by the IGN. Its objective is to correct the toponymy of this national gazetteer according to the documentation produced by the autonomous community entities, which are generally competent in the matter. The methodology, called "autocorrection", was designed around 2011 and has had different degrees of development. The evaluation of the pilot projects, implemented in the provinces of Huelva and Alava, shows a degree of exact coincidence between databases of 30% [17], although certain problems are evident in cases of bilingualism, such as an increase in errors and spelling variations. The manual component of validation of all correspondences has a great weight.

On the other hand, this implementation has brought homogeneity and reliability to the NGBE, in addition to promoting a vigorous activity of the administration with the aim of complying with the requirements derived from the European INSPIRE directive. It also brings an essential idea, which is that the improvement of these databases can only be approached iteratively. Whereas the spirit of the project may be aligned with this work, the results are hardly comparable, and there is a lack of descriptive statistics for the geographic scope of our study.

The methodology design is based on GIS tools, which provide all the power of spatial analysis. On the other hand, the capacity of ETL tools has been used for managing databases and their transformation based on alphanumeric data. Although the method can be fully reproduced by database managers or by GIS, there are several advantages of ETL tools. Firstly, there is the agility of data loading (the cadastre's working unit is the municipality, so there are multiple files to be linked); then the operability between formats; and finally, the possibility of implementing subroutines with code adapted to the needs of the data models used. The difference between data models depending on the territorial unit analyzed is one of the main challenges of the extension of the application of this method to other datasets with toponymic content.

Nevertheless, different projects are known to offer solutions to similar problems, so it is necessary to discuss the choice of the tools used. Let us consider the design implemented by Kaufman [102], which searches for matches using an algorithmic solution based on the fuzzy string matching methodology. We find that our proposal, which uses matching with a geographic LKB, is better adapted to the Basque language construction logic of lexeme agglutination. The use of natural language processing tools using artificial intelligence [106] should also be considered. Nevertheless, this route requires a deeper immersion into the linguistic concepts beyond the objectives of the line of research in which this study is embedded.

Throughout this manuscript, we have also indicated the different limitations in the scope of the work. Due to the initial geographic perspective, we have not considered the implementation of forms of natural language processing using artificial intelligence, which has yielded very favorable results, as mentioned in Sections 2.2 and 2.4. In addition, a toponymic standardization process has yet to be deployed on cadastral geographic information, which makes it particularly difficult to compare with linguistically standardized databases. On the other hand, this work leaves aside the typification of geographic elements. Thus, this could be one of the objectives of future work on the corpus created with the set of place names. Finally, it must be admitted that any systematic toponymy treatment offers improvements in homogeneity and speed in obtaining results. Nevertheless, it needs to improve control in the management of detail and the interpretation of nuances, which is possible in manual methods.

Concerning the semantic aspect, it is possible to outline an action to improve the procedure by including it in the reference LKB terminology related to botany, land use, and land cover, in addition to geographical terms. Therefore, it would be advisable to design a semantic map of the organization of the lexical system related to toponymy [107], which would endow a semantic affiliation to different categories in order to facilitate the application of this LKB.

7. Conclusions

In this research, we have designed, implemented, and validated a methodology for the detection of coincidences between geographic databases using ETL and GIS tools. These have been the NGO-CAE gazetteer and the cadastral geoinformation corresponding to the province of Gipuzkoa (bilingual territorial). The objective of this comparison is the integration of information at an academic level. In addition, toponymy is a special type of geoinformation because of its transversality and the multitude of approaches from which it can be addressed.

The implemented methodology is based on spatial criteria, textual element treatments, and finally on semantic criteria, which outline the automatic selection of concordances, as verified in the manual validation of the method, performing the same process manually. The validation phase of the method on a sample of four municipalities clarifies the minimum concordance chain length in which semantic units are found. These are crucial data to determine the threshold at which the processing is reliable by exclusively automatic methods. In our case, as Basque is the predominant language, it is found that strings of fewer than four characters cease to have a unique significance, and indeterminacy takes over the analysis so that the execution of the succession of commands is not enough. It has also been found that the more or less restrictive use of LKB leads to error modulation. For the issue that concerns us, which is the integration of databases, the errors of commission result in the potential loss of toponymic information. Thereby, it will be the parameter that we must consider to limit the terminological glossary.

The results achieved in this implementation, considering the high number of non-repeated names, demonstrate this method's goodness and the adequacy of the choice of databases for their integration. A lay view might think this low consistency may denote a lack of quality in the compared sources. However, given the nature of the sources, the interpretation derived from the detailed knowledge of the datasets tends to

relate this incoherence to differences in the times and methods of information capture, or types of entities consigned for the purpose of each work. In short, this work confronts different approaches, objectives, and requirements of each representation of reality through geographic information.

If this study is contrasted with other above-mentioned related works, we can assert that the goodness of the results is comparable. Furthermore, this example provides an innovative solution for the treatment of toponymy in a bilingual environment. Finally, it explores the toponymic possibilities of a geoinformation database, such as that of the Gipuzkoa land registry, which until now has not been used for these goals.

The assumption of the challenge of extending the geographical scope of application to the other territories of the Basque Country, Alava, or Bizkaia must involve result verification due to the population and sociolinguistic differences. Finally, and from a more general perspective, the main challenge is to advance toward automated, systematic, and reliable management in a field of knowledge that has been treated by hand until very recently.

Author Contributions: Conceptualization, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; methodology, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; software, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; validation, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; formal analysis, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; investigation, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; resources, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; data curation, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; writing—original draft preparation, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; writing—review and editing, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; visualization, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; supervision, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; project administration, Oihana Mitxelena-Hoyos and José-Lázaro Amaro-Mellado; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

```
#Subroutine input variables:
#CsvNombreC: cadastral parcel toponym
#Relationships{}.NOMBRET: dictionary that contains the place names that fulfil the geometric criterion (see Section 4.3.3. Geometric Criterion)
#ELEMENTOS_EN_LISTA: number of elements in Relationships{}:
#Subroutine output variables:
#For every feature of Relationships{} dictionary, we have the following variables:
#Relationships{}.COMMON_CHAIN: the maximum-common-chain text.
#Relationships{}.LEN_ COMMON_CHAIN: length of the maximum-common-chain text
import fme
import fmeobjects

def FeatureProcessor(feature):

        nombre = str(feature.getAttribute('csvNombreC'))
        m = len(nombre)
        num_ele_list = feature.getAttribute('ELEMENTOS EN LISTA')
        count=0
```

```
while count < num_ele_list:
    toponimo = feature.getAttribute('_relationships{' + str(count) + '}.NOMBRET')
    n = len(str(toponimo))
    maxLen = 0
    endIndex = m
    FIND = [[0 for x in range(n + 1)] for y in range(m + 1)]

    for i in range(1, m + 1):
        for j in range(1, n + 1):

            if nombre[i − 1] == toponimo[j − 1]:
                FIND[i][j] = FIND[i − 1][j − 1] + 1

                if FIND[i][j] > maxLen:
                    maxLen = FIND[i][j]
                    endIndex = i

    resultado = nombre[endIndex − maxLen: endIndex]

    feature.setAttribute('_relationships{' + str(count) + '}.COMMON_CHAIN',str(resultado))
    feature.setAttribute('_relationships{' + str(count) + '}.LEN_COMMON_CHAIN',len
(str(resultado)))
    count = count + 1
```

References

1. Quesada-García, S. A cartography of al-Andalus' landscape: Mapping settlements of Muslim agricultural colonization in Europe applying GIS techniques. *J. Hist. Geogr.* **2022**, *77*, 65–84. [CrossRef]
2. Rosselló i Verger, V.M. Cartography, landscape and territory. *Catalan Soc. Sci. Rev.* **2012**, *1*, 46–57. [CrossRef]
3. Zamorshchikova, L.; Gadal, S.; Filippova, V.; Samsonova, M. Landscape Toponymic Maps: Interdisciplinary Approach (Example of Sakha Republic, Russia). In Proceedings of the International Multidisciplinary Scientific Conference SGEM2016, Albena, Bulgaria, 30 June–6 July 2016; pp. 311–316.
4. Arroyo Illera, F. Toponymy as a intangible cultural legacy. *Boletín Real Soc. Geogr.* **2019**, *153*, 33–60.
5. Mollo, N.M. Determinación geográfica de los sitios de interés histórico y arqueológico mediante la utilización de técnicas cartográficas. *Teor. Práct. Arqueol. Hist. Latinoam.* **2022**, *3*, 21–48. [CrossRef]
6. Ingelmo Casado, R. Localización y tratamiento de información histórica a través de la toponimia menor: Utilidad del catastro de la riqueza rústica (Localization and treatment of historical information through the minor toponymy: Utility of the cadastre of rustic wealth). In *Tecnologías de la Información Geográfica: La Información Geográfica al Servicio de los Ciudadanos (Geographic Information Technologies: Geographic Information at the Service of Citizens)*; Ojeda, J., Piya, M.F., Vallejo, I., Eds.; Secretariado de Publicaciones de la Universidad de Sevilla: Sevilla, Spain, 2010; pp. 199–213, ISBN 978-84-472-1294-1.
7. Mácha, P.; Obrusník, U.; Jordan, P.; Sancho Reinoso, A. The Challenges of Studying Place-Name Politics in Multilingual Areas. In *Place-Name Politics in Multilingual Areas*; Springer International Publishing: Cham, Switzerland, 2021; pp. 45–69.
8. González García, E.M. El catastro: Fuente de información del territorio (The cadastre: Source of territorial information). In *Proceedings of the X Coloquio de Historia Canario—Americano. Coloquio 10. Tomo 2*; Cabildo Insular de Gran Canaria, Ed.; ULPGC: Las Palmas de Gran Canaria, Spain, 1992; pp. 160–175.
9. European Parliament and Council of the European Union. *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infraestructure for Spatial Information in the European Community (INSPIRE), L108*; Official Journal of the European Union: Brussels, Belgium, 2007.
10. Gobierno de España. *Ley 14/2010, de 5 de Julio, sobre las Infraestructuras y los Servicios de Información Geográfica en España (LISIGE) (Law 14/2010, of July 5, 2010, on Geographic Information Infrastructures and Services in Spain)*; Government of Spain: Madrid, Spain, 2010.
11. Liu, Z.; Cheng, L. Review of GIS Technology and Its Applications in Different Areas. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *735*, 012066. [CrossRef]
12. Mesquitela, J.; Elvas, L.B.; Ferreira, J.C.; Nunes, L. Data Analytics Process over Road Accidents Data—A Case Study of Lisbon City. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 143. [CrossRef]
13. Páez, O.; Vilches-Blázquez, L.M. Bringing Federated Semantic Queries to the GIS-Based Scenario. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 86. [CrossRef]
14. Drešček, U.; Kosmatin Fras, M.; Tekavec, J.; Lisec, A. Spatial ETL for 3D Building Modelling Based on Unmanned Aerial Vehicle Data in Semi-Urban Areas. *Remote Sens.* **2020**, *12*, 1972. [CrossRef]

15. García-Balboa, J.; Ureña-Cámara, M.; Ariza-López, F.; Garrido-Borrego, M.T.; Torrecillas-Lozano, C. Análisis comparativo entre la base de datos de toponimia 1:10,000 (BTA10) y la toponimia contenida en la base de datos catastral (Comparative analysis between the toponymy database 1:10,000 (BTA10) and the toponymy contained in the cadastral database). In Proceedings of the I Congreso Internacional sobre Catastro Unificado Multipropósito (CICUM), Jaén, Spain, 16–18 June 2010; Alcázar-Molina, M.G., Ariza-López, F.J., García-Balboa, J.L., Mata-de-Castro, E., Ruiz-Capiscol, S., Ureña-Cámara, M.A., Eds.; Universidad de Jaén: Jaén, Spain, 2010; pp. 271–277.
16. Garrido, M.; Torrecillas, C. Gestión toponímica vinculada a la cartografía en Andalucía (España) (Toponymic management linked to cartography in Andalusia (Spain)). *Rev. Bras. Geogr.* **2022**, *67*, 120.
17. Vázquez Hoehne, A.; Rodríguez de Castro, A.; Luján Díaz, A.; Montilla Lillo, M.; Castaño Suárez, A. Propuesta metodológica para la elaboración del Nomenclator Geográfico Básico de España a apartir de la autocorrección de la Base Cartográfica Numérica con la información de las comunidades autómomas (Methodological proposal for the elaboration of the Basic Geographic Nomenclator of Spain from the autocorrection of the Numerical Cartographic Base with the information of the autonomous communities). In *Proceedings of the Els noms en la vida quotidiana. Actes del XXIV Congrés Internacional d'ICOS sobre Ciències Onomàstiques. Annex. Secció 11*; Onomástica; Biblioteca Técnica de Política Lingüística: Barcelona, Spain, 2011.
18. Köhnlein, B. The morphological structure of complex place names: The case of Dutch. *J. Comp. Ger. Linguist.* **2015**, *18*, 183–212. [CrossRef]
19. Rodríguez Pérez, C.; Castañón Álvarez, J.C. Modos de representación cartográfica de las unidades de paisaje: Revisión y propuestas (Modes of cartographic representation of landscape units: Review and proposals). *Ería* **2016**, *99*, 15–40. [CrossRef]
20. Garrido Villén, N. Propuesta de Normalización Cartográfica para el Desarrollo Territorial. Ph.D. Thesis, Universitat Politècnica de València, Valencia, Spain, 2008.
21. Mango, J.; Claramunt, C.; Ngondo, J.; Zhang, D.; Xu, D.; Colak, E.; Li, X. Multipurpose temporal GIS model for cadastral data management. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 1205–1230. [CrossRef]
22. Frajer, J.; Fiedor, D. Discovering extinct water bodies in the landscape of Central Europe using toponymic GIS. *Morav. Geogr. Rep.* **2018**, *26*, 121–134. [CrossRef]
23. Gordova, Y.; Herzen, O.; Herzen, A.; Kostovska, S. Usage of cartographic methods in place-name study (history of the problem and actual research). *InterCarto. InterGIS* **2021**, *27*, 520–536. [CrossRef]
24. Rodríguez de Castro, A.; Vázquez Hoehne, A. Decoding of Place Names as Geographical Information Tools. *Mitt. Osterr. Geogr. Ges.* **2019**, *158*, 263–288. [CrossRef]
25. Giraut, F.; Houssay-Holzschuch, M. Place Naming as *Dispositif*: Toward a Theoretical Framework. *Geopolitics* **2016**, *21*, 1–21. [CrossRef]
26. Bijak, U. Space and Landscape in Polish Toponymy. *J. Linguist. Cas.* **2021**, *72*, 194–207. [CrossRef]
27. Atik, M.; Kanabakan, A.; Ortaçeşme, V.; Yildirim, E. Tracing landscape characters through place names in rural Mediterranean. *CATENA* **2022**, *210*, 105912. [CrossRef]
28. Felecan, O. Romanian Oikonyms and Hodonyms Mirroring the Great Union of 1918. *Mitt. Osterr. Geogr. Ges.* **2021**, *1*, 495–517. [CrossRef]
29. Membrado-Tena, J.C. El papel de la Geografía en el análisis del contenido semántico de los topónimos. El caso de Alicante. *An. Geogr. Univ. Complut.* **2018**, *38*, 35–60. [CrossRef]
30. Liao, Y.-P.; Lin, F.-T. Place Name Ambiguities in Urban Planning Domain Ontology. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 12–14 November 2015; SCITEPRESS—Science and and Technology Publications: Setúbal, Portugal, 2015; pp. 429–434.
31. Membrado-Tena, J.C.; Fansa, G. Toponimia, paisaje y ciencia. El caso de los nombres de municipio de la Plana de Castelló (País Valenciano). *Cuad. Geográficos* **2020**, *59*, 28–52. [CrossRef]
32. Maus, G. Landscapes of memory: A practice theory approach to geographies of memory. *Geogr. Helv.* **2015**, *70*, 215–223. [CrossRef]
33. Mulrennan, M.E. Do Landscapes Listen? Wemindji Eeyou Knowledge, Adaptation and Agency in the Context of Coastal Landscape Change. In *Landscapes and Landforms of Eastern Canada*; Springer: Cham, Switzerland, 2020; pp. 543–556.
34. De Felice, P.; La Greca, F.; Siniscalchi, S. The place names of the Aragonese maps. Interpretative hypotheses, landscape readings and methodological proposals. *Int. J. Cartogr.* **2022**, *8*, 3–17. [CrossRef]
35. Bahgat, K.; Runfola, D. Toponym-assisted map georeferencing: Evaluating the use of toponyms for the digitization of map collections. *PLoS ONE* **2021**, *16*, e0260039. [CrossRef] [PubMed]
36. Rowland, A.; Folmer, E.; Beek, W. Towards Self-Service GIS—Combining the Best of the Semantic Web and Web GIS. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 753. [CrossRef]
37. Wang, L. Research on the design of large data storage structure of database based on Data Mining. In Proceedings of the 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture, Manchester, UK, 23–25 October 2021; ACM: New York, NY, USA, 2021; pp. 3001–3004.
38. Nys, G.-A.; Dubois, C.; Goffin, C.; Hallot, P.; Kasprzyk, J.-P.; Treffer, M.; Billen, R. Geodata quality assessment and operationalisation of the INSPIRE directive: Feedback. *Bull. Soc. Géogr. Liège* **2022**, *78*, 179–188. [CrossRef]
39. Conedera, M.; Vassere, S.; Neff, C.; Meurer, M.; Krebs, P. Using toponymy to reconstruct past land use: A case study of 'brüsáda' (burn) in southern Switzerland. *J. Hist. Geogr.* **2007**, *33*, 729–748. [CrossRef]
40. Pijet-Migoń, E.; Migoń, P. Geoheritage and Cultural Heritage—A Review of Recurrent and Interlinked Themes. *Geosciences* **2022**, *12*, 98. [CrossRef]
41. Blaschke, T.; Merschdorf, H.; Cabrera-Barona, P.; Gao, S.; Papadakis, E.; Kovacs-Györi, A. Place versus Space: From Points, Lines and Polygons in GIS to Place-Based Representations Reflecting Language and Culture. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 452. [CrossRef]

42. Serikova, S.; Baishukurova, G. Digital technologies in the toponymy study. *SHS Web Conf.* **2022**, *141*, 04001. [CrossRef]
43. Hernández, F.; Rosales, J.; Rodrigo, J. Búsquedas inteligentes de toponimia (Intelligent toponymy searches). In Proceedings of the V Jornadas Técnicas de la IDE de España, JIDEE 2008, Santa Cruz de Tenerife, Spain, 5–7 November 2008.
44. Gordón-Peral, M. *Toponimia de España. Estado Actual y Perspectivas de la Investigación (Toponymy of Spain. Current Status and Research Perspectives)*; Gordón-Peral, M.D., Ed.; Walter de Gruyter: Berlin, Germany, 2010; Volume 24, ISBN 978-3-11-023348-3.
45. Grossner, K.; Grunewald, S.; Mostern, R. Bringing places from the distant past to the present: A report on the World Historical Gazetteer. *Int. J. Digit. Libr.* **2022**, *Volume*, Page. [CrossRef]
46. Garrido Borrego, M.T.; Torrecillas, C.; Benabad, I.G.; Cardenas, L.R.; Nicolás, C.T. Standardization of geospatial data of water sources and springs collected in the Andalusian Gazetteer (Spain). *Rev. Cart.* **2021**, *103*, 99–121. [CrossRef]
47. Gelernter, J.; Ganesh, G.; Krishnakumar, H.; Zhang, W. Automatic gazetteer enrichment with user-geocoded data. In Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Orlando, LA, USA, 5 November 2013; ACM: New York, NY, USA, 2013; pp. 87–94.
48. Acheson, E.; Volpi, M.; Purves, R.S. Machine learning for cross-gazetteer matching of natural features. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 708–734. [CrossRef]
49. Hearn, K.P. Mapping the past: Using ethnography and local spatial knowledge to characterize the Duero River borderlands landscape. *J. Rural Stud.* **2021**, *82*, 37–53. [CrossRef]
50. Hagemans, E.; Unger, E.-M.; Soffers, P.; Wortel, T.; Lemmen, C. The new, LADM inspired, data model of the Dutch cadastral map. *Land Use Policy* **2022**, *117*, 106074. [CrossRef]
51. Das, S.N.; Sreenivasan, G.; Srinivasa Rao, S.; Joshi, A.K.; Varghese, A.O.; Prakasa Rao, D.S.; Chandrasekar, K.; Jha, C.S. Geospatial Technologies for Development of Cadastral Information System and its Applications for Developmental Planning and e-Governance. In *Geospatial Technologies for Resources Planning and Management*; Springer: Cham, Switzerland, 2022; pp. 485–538.
52. Vilches-Blázquez, L.M.; Saavedra, J. A graph-based representation of knowledge for managing land administration data from distributed agencies—A case study of Colombia. *Geo-Spat. Inf. Sci.* **2022**, *25*, 259–277. [CrossRef]
53. Atazadeh, B.; Olfat, H.; Rajabifard, A.; Kalantari, M.; Shojaei, D.; Marjani, A.M. Linking Land Administration Domain Model and BIM environment for 3D digital cadastre in multi-storey buildings. *Land Use Policy* **2021**, *104*, 105367. [CrossRef]
54. Gobierno de España. *Ley 13/2015, de 24 de Junio, de Reforma de la Ley Hipotecaria (Law 13/2015, of June 24, 2015, on the Reform of the Mortgage Law)*; Government of Spain: Madrid, Spain, 2015.
55. Velilla-Torres, J.M.; Mora-Navarro, G.; Femenia-Ribera, C.; Martínez-Llario, J.C. The georreferenced alternative graphic representation in the Cadastre-Land Registry coordination in Spain. Implementation study of the ISO-19152 (LADM) at the international level. In Proceedings of the 1st Congress in Geomatics Engineering—CIGeo, Valencia, Spain, 5–6 July 2017; Universitat Politècnica València: Valencia, Spain, 2017; pp. 69–76.
56. Krigsholm, P.; Riekkinen, K.; Ståhle, P. The Changing Uses of Cadastral Information: A User-Driven Case Study. *Land* **2018**, *7*, 83. [CrossRef]
57. Atazadeh, B.; Halalkhor Mirkalaei, L.; Olfat, H.; Rajabifard, A.; Shojaei, D. Integration of cadastral survey data into building information models. *Geo-Spat. Inf. Sci.* **2021**, *24*, 387–402. [CrossRef]
58. Femenia-Ribera, C.; Mora-Navarro, G.; Pérez, L.J.S. Evaluating the use of old cadastral maps. *Land Use Policy* **2022**, *114*, 105984. [CrossRef]
59. Duque, J.I.G. El Catastro de Rústica de Guipúzcoa/Gipuzkoa. *Foresta* **2012**, *55*, 180–185.
60. Gordón-Peral, M. Las fuentes de documentación toponímica: El catastro del Marqués de la Ensenada y su interés lingüístico (The sources of toponymic documentation: The cadastre of the Marqués de la Ensenada and its linguistic interest). In *Indagaciones Sobre la Lengua: Estudios de Filología y Lingüística Españolas en Memoria de Emilio Alarcos*; University of Seville: Seville, Spain, 2001; pp. 437–454, ISBN 84-472-0682-3.
61. Houssay-Holzschuch, M.; Giraut, F. *Politics of Place Naming Naming the World*; Wiley & Sons: Hoboken, NJ, USA, 2022.
62. Sanz Elorza, M.; González Bernardo, F. Toponimia de origen vegetal en la provincia de Segovia a partir de los datos del Catastro de Rústica (Toponymy of vegetal origin in the province of Segovia based on the data of the Rural Cadastre). *Lazaroa* **2006**, *27*, 103–125.
63. Ingelmo Casado, R. Georreferenciación de documentación histórica mediante la toponimia de los catastros (Georeferencing of historical documentation using the toponymy of land registry records). *Int. Rev. Geogr. Inf. Sci. Technol. Geofocus* **2012**, *12*, 243–267.
64. Skorup, D. Models of utility cadastre in Bosnia and Herzegovina. *AGG+ J. Archit. Civ. Eng. Geod. Other Relat. Sci. Fields* **2020**, *8*, 77–88. [CrossRef]
65. Sánchez Ondoño, I.; Cebrián Abellán, F.; Garcia-Gonzalez, J.A. The Cadastre as a Source for the Analysis of Urbanization Dynamics. Applications in Urban Areas of Medium-Sized Inland Spanish Cities. *Land* **2021**, *10*, 374. [CrossRef]
66. Pearn, J. Surveyors, toponymy and heritage. *Queensl. Hist. J.* **2022**, *25*, 175–188.
67. Cienciała, A.; Sobolewska-Mikulska, K.; Sobura, S. Credibility of the cadastral data on land use and the methodology for their verification and update. *Land Use Policy* **2021**, *102*, 105204. [CrossRef]
68. Gobierno de Aragón (Spain). *Nomenclátor Geográfico de Aragón*; Government of Aragon: Zaragoza, Spain, 2019.
69. Real Academia Galega. Xunta de Galicia (Spain) Toponimia de Galicia. Available online: https://toponimia.xunta.gal/es/toponimia (accessed on 4 February 2023).

70. Fons, M.; Gomilla, X. Sobre la situación de la toponimia oficial en las Illes Balears: El Nomenclátor de Toponimia de Menorca y el futuro Nomenclátor Geográfico de las Illes Balears (On the situation of the official toponymy in the Balearic Islands: The Gazetteer of Toponymy of Menorca and the future Geographic Gazetteer of the Balearic Islands). *Mapping* **2019**, *28*, 48–56.
71. Zubiaur-Carreño, F. Toponimia de San Martín de Unx según los amojonamientos de villa en el siglo XVI (Toponymy of San Martín de Unx according to the sixteenth-century town demarcations). *Cuad. Etnol. Etnogr. Navar.* **1978**, *29*, 255–272.
72. Etxebarria, M. *Principios y Fundamentos de Sociolingüística (Principles and Fundamentals of Sociolinguistics)*; UPV/EHU, Ed.; University of the Basque Country: Leioa, Spain, 2000; Volume 488.
73. Benito del Valle Eskauriaza, A. Linguistic Trends of Basque Youth. Minorization and Visibilization in the Spanish and European Context. *Cuad. Eur. Deusto* **2022**, *4*, 107–135. [CrossRef]
74. Jordan, P.; Mácha, P.; Balode, M.; Krtička, L.; Obrusník, U.; Pilch, P.; Sancho Reinoso, A. *Place-Name Politics in Multilingual Areas*; Springer International Publishing: Cham, Switzerland, 2021; ISBN 978-3-030-69487-6.
75. Ruiz Vieytez, E.J. The Future of European Minority Languages: A Normative Analysis. *Cuad. Eur. Deusto* **2022**, *4*, 37–67. [CrossRef]
76. Sanchez, J.M. Bilingüismo, disglosia, contacto de lenguas (Bilingualism, dysglossia, language contact). In *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*; UPV/EHU University of the Basque Country: Bilbao, Spain, 1974; Volume 8, pp. 3–79.
77. Brinkmann, L.M.; Duarte, J.; Melo-Pfeifer, S. Promoting Plurilingualism Through Linguistic Landscapes: A Multi-Method and Multisite Study in Germany and the Netherlands. *TESL Can. J.* **2022**, *38*, 88–112. [CrossRef]
78. Marten, H.F.; Van Mensel, L.; Gorter, D. Studying Minority Languages in the Linguistic Landscape. In *Minority Languages in the Linguistic Landscape*; Palgrave Macmillan UK: London, UK, 2012; pp. 1–15.
79. Hallett, R.W.; Quiñones, F.M. The linguistic landscape of an Urban Hispanic-Serving Institution in the United States. *Soc. Semiot.* **2021**, *7*, 1–15. [CrossRef]
80. Rugkhapan, N.T. Between Toponymy and Cartography: An Evolving Geography of Heritage in George Town, Malaysia. In *New Directions in Linguistic Geography*; Springer Nature Singapore: Singapore, 2022; pp. 327–354.
81. De Viñaspre, R.G. Euskara batuaren sorrera eta ibilbidea: Euskara batua eta toponimia (The creation and trajectory of the united Basque: The united Basque and toponymy). In *Arantzazutik Mundu Zabalera: 1968–2018 = La Normativización del Euskera: 1968–2018 = La Standardisation de la Langue Basque: 1918–2018 = Standardization of the Basque Language: 1968–2018*; Iberoamericana Vervuert: Madrid, Spain, 2022; pp. 241–252.
82. Salaberri, P. Arauketa eta ikerketa, elkarren adiskide beharrak (Regulation and research, mutual needs). In *Euskal Onomastika Aplikatua XXI. Mendean*; Iberoamericana Editorial Vervuert: Madrid, Spain, 2020; Volume 39, pp. 155–168.
83. JGG Juntas Generales de Gipuzkoa. Available online: https://www.bngipuzkoa.eus/ (accessed on 27 November 2022).
84. Álvarez Pastor, M.; Ubillos Pernaut, J.; Muniategiandikoetxea Markiegi, J.; Valenciano Tamayo, T.; García Odiaga, I. Los límites fronterizos como aglutinadores de actividad en el territorio La frontera del Bidasoa (Border limits as agglutinators of activity in the territory The Bidasoa Border). In Proceedings of the Actas de los Seminarios de Apoyo a la Investigación hibridación y Transculturalidad en los modos de Habitación Contemporánea; Tapia Martín, C., Varona Gandulfo, M., Eds.; University of Seville: Seville, Spain, 2009; pp. 417–424.
85. Lozano Valencia, M.A.; Lozano Valencia, P.J. Service mancommunities: An example of territorial vertebration in Guipuzcoa. A description of domestic waste in the territory. *Bol. Asoc. Geógr. Esp.* **2008**, *48*, 417–420.
86. Dávila-Cabanillas, N. The rural enviroment in the territorial planning of the CAPV. *Lurralde* **2022**, *45*, 1–32.
87. Membrado-Tena, J.C. Place names as indicators of extinct landscapes. The case of Castelló de la Plana. *Cuad. Geogr.* **2022**, *108*, 461–480.
88. Cortes Generales (Spain). *Constitución Española*; Cortes Generales (General Courts): Madrid, Spain, 1978.
89. Gobierno Vasco (Spain). *V Mapa Sociolingüistico, 2011*, 1st ed.; Servicio de Publicaciones del Gobierno Vasco: Vitoria-Gasteiz, Spain, 2014.
90. Trapero, M. *Para una Teoría Lingüística de la Toponimia Estudios de Toponimia Canaria*; ULPGC. Servicio de Publicaciones y Difusión Científica: Las Palmas de Gran Canaria, Spain, 1995; Volume 3.
91. Gobierno-Vasco-España. *Decreto 179/2019, de 19 de Noviembre, sobre Normalización del uso Institucional y Administrativo de las Lenguas Oficiales en las Instituciones Locales de Euskadi (Decree 179/2019, of November 19, on the Standardization of the Institutional and Administrative Use of the Official Languages in the Local Institutions of the Basque Country)*; Basque Government: Vitoria-Gasteiz, Spain, 2019.
92. Council-of-Europe. *European Charter for Regional or Minority Languages (ETS No. 148)*; Council-of-Europe: Strasbourg, France, 1992.
93. Jefatura-del-Estado-España. *Instrumento de Ratificación de la Carta Europea de las Lenguas Regionales o Minoritarias, Hecha en Estrasburgo el 5 de Noviembre de 1992 (Instrument of Ratification of the European Charter for Regional or Minority Languages, Done at Strasbourg on November 5, 1992)*; Head of State-Spain: Madrid, Spain, 2001.
94. Departamento de Cultura y Política Lingüística Nomenclator Geográfico Oficial de la CAE. Available online: https://www.euskadi.eus/toponima-onomastica-cav/web01-a2corpus/es/ (accessed on 28 July 2022).
95. Gobierno de España. *Real Decreto 1545/2007, de 23 de Noviembre, por el que se Regula el Sistema Cartográfico Nacional (Royal Decree 1545/2007, of November 23, 2007, which regulates the National Cartographic System)*; Government of Spain: Madrid, Spain, 2007.
96. Figurska, M.; Dawidowicz, A.; Zysk, E. Voronoi Diagrams for Senior-Friendly Cities. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7447. [CrossRef]
97. LaForce, T.; Ebeida, M.; Jordan, S.; Miller, T.A.; Stauffer, P.H.; Park, H.; Leone, R.; Hammond, G. Voronoi Meshing to Accurately Capture Geological Structure in Subsurface Simulations. *Math. Geosci.* **2022**, *55*, 129–161. [CrossRef]

98. Qi, Y.; Wang, R.; He, B.; Lu, F.; Xu, Y. Compact and Efficient Topological Mapping for Large-Scale Environment with Pruned Voronoi Diagram. *Drones* **2022**, *6*, 183. [CrossRef]
99. Dalvi, N.; Olteanu, M.; Raghavan, M.; Bohannon, P. Deduplicating a places database. In Proceedings of the 23rd international conference on World wide web—WWW '14, Seoul, Republic of Korea, 7–11 April 2014; ACM Press: New York, NY, USA, 2014; pp. 409–418.
100. Instituto Geográfico Nacional. *MTN25. Normas de Edición 1:25000 (1:25000 Edition Standards)*; Instituto Geográfico Nacional: Madrid, Spain, 2014.
101. Black, P.E. "Longest Common Substring" in Dictionary of Algorithms and Data Structures. Available online: https://www.nist.gov/dads/HTML/longestCommonSubstring.html (accessed on 27 September 2022).
102. Kaufman, A.R.; Klevs, A. Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field. *Polit. Anal.* **2022**, *30*, 590–596. [CrossRef]
103. Pociello, E.; Agirre, E.; Aldezabal, I. Methodology and construction of the Basque WordNet. *Lang. Resour. Eval.* **2011**, *45*, 121–142. [CrossRef]
104. Azcárate Luxán, M.; Alcázar González, A.; García Cancela, X.; Gorrotxategi Nieto, M.; Vives Pérez i Piquer, A. *Toponymic Guidelines for Map and Other Editors for International Use (Spain)*; Azcárate Luxán, M., Ed.; Centro Nacional de Información Geográfica: Madrid, Spain, 2012.
105. Ugarte Garrido, J. II Congreso geoEuskadi (2nd geoEuskadi Congress). In *Proceedings of the Toponimia eta EAEko Izendegi Geografiko Ofiziala*; Departamento de Planificación Territorial, Vivienda y Transportes: Bilbao, Spain, 2021.
106. Blandón Andrade, J.C. Applications of natural language processing. In *Entre Ciencia e Ingeniería*; Universidad Católica de Pereira: Pereira, Colombia, 2022; Volume 16, pp. 7–8.
107. Rakhilina, E.; Ryzhova, D.; Badryzlova, Y. Lexical typology and semantic maps: Perspectives and challenges. *Z. Sprachwiss.* **2022**, *41*, 231–262. [CrossRef]