Approximating Spatial Distance Through Confront Networks: Application to the Segmentation of Medieval Avignon

Margot Ferrand, Vincent Labatut

Approximating Spatial Distance Through *Confront* Networks: Application to the Segmentation of Medieval Avignon

Margot Ferrand[1] and Vincent Labatut[2]

[1]CIHAM – UMR 5648, 74 rue Louis-Pasteur, Avignon, F-84029, France
Ville d'Avignon, Maison du Patrimoine, 20 rue du Roi René, Avignon, F-84045, France. email: margot.ferrand@alumni.univ-avignon.fr
[2]LIA – UPR 4128, Avignon Université, 339 chemin des Meinajariès, Avignon, F-84911, France. email: vincent.labatut@univ-avignon.fr

November 19, 2024

Abstract

In historical studies, the older the sources, the more common it is to have access to data that are only partial, and/or unreliable or imprecise. This can make it difficult, or even impossible, to perform certain tasks of interest, such as the segmentation of some urban space based on the location of its constituting elements. Indeed, traditional approaches to tackle this specific task require knowing the position of all these elements before clustering them. Yet, alternative information is sometimes available, which can be leveraged to address this challenge. For instance, in the Middle Ages, land registries typically do not provide exact addresses, but rather locate spatial objects relative to each other, e.g. x being to the North of y. Spatial graphs are particularly adapted to model such spatial relationships, called *confronts*, which is why we propose their use over standard tabular databases. However, historical data are rich and allow extracting confront networks in many ways, making the process non-trivial. In this article, we propose several extraction methods and compare them to identify the most appropriate. We postulate that the best candidate must constitute an optimal trade-off between covering as much of the original data as possible, and providing the best graph-based approximation of spatial distance. Leveraging a dataset that describes Avignon during its papal period, we show empirically that the best results require ignoring some of the information present in the original historical sources, and that including additional information from secondary sources significantly improves the confront network. We illustrate the relevance of our method by partitioning the best graph that we extracted, and discussing its community structure in terms of urban space organization, from a historical perspective. Our data and source code are both publicly available online.

Keywords— Graph Extraction, Medieval History, Land Registries, Spatial Networks, Community Detection

Cite as: Margot Ferrand & Vincent Labatut, Approximating Spatial Distance Through Confront Networks: Application to the Segmentation of Medieval Avignon, *Journal of Complex Networks*, 2024, in press.

1 Introduction

To study urban spatiality in medieval history, it is necessary to cross-reference different types of sources related to the morphology and use of space: written, planimetric, iconographic, and archaeological sources [1]. The relevant written sources are particularly numerous and can be of various kinds: legal, tax or religious records, census data, chronicles and annals, correspondence, testaments... Among them, land documentation, which encompasses inventories of properties, administrative surveys, terriers, censiers, and recognition books, is particularly useful when focusing on the Middle Age, due to the state of medieval archives. Indeed, from the 13th century onwards, this documentation increases considerably and constitutes a major part of the written records [3]. Regardless of the language in which they are written, these documents have common characteristics. Each one of them contains a series of declarations in which tenants declare the properties they hold under the direct domain of a lord and for which they must pay an annual fee.

However, historians encounter serious difficulties in dealing with land documentation, which is too often confined to strictly accounting treatment, thereby overlooking what these sources can reveal in terms of the uses and representation of urban space. The thorny issue with this documentation is the spatialization of information, to which its incompleteness must often be added. Indeed, historians who have studied this documentation have often tried to reconstruct medieval parcels, i.e. to locate them as precisely as possible on a map [18]. But this task is almost impossible: before the modern era, it is very rare for these sources to be accompanied by maps, and the provided locations do not yet refer to precise addresses. They are only relative, and difficult to interpret. In this interdisciplinary work combining History, Geography, Computer Science, and Network Science, we show that this step is not mandatory to analyze the distribution of properties recorded in the land documentation and to fully exploit these sources. Historical analysis at the city scale does not need the land holdings to be precisely located but rather requires that relative locations be possible, in order to interpret the distribution of masses, the filled and empty spaces, proximities and distances, attractions, and separations.

In this article, we propose a graph-based approach to model the urban space based on historical sources. Its main advantage is to focus on the spatial *relationships* between the represented objects, e.g. x being to the North of y. By using such a *relative* way of locating these objects, our method does not require estimating their *absolute* location on a map, which is a major issue of the standard approach. As a consequence, our method can handle incomplete data that would provide the absolute location of only a part of the described objects. The resulting graph is designed to encode the notion of spatial proximity, and can be used as an approximate representation of the urban space to conduct tasks such as spatial segmentation.

Our contributions are two-fold. On the methodological side, we propose several methods to extract so-called *confront* networks from historical data taking the form of land registries. We also propose two objective criteria to compare these extraction methods and identify the most appropriate one. On the applications side, we present a dataset describing the city of Avignon during the papacy, and use it as a benchmark for our methods. We extract and compare 12 different versions of the confront network, and select the most suitable to a spatial analysis. We then show empirically its relevance by segmenting the urban space through community detection, and discussing the historical value of the resulting subdivisions of the city.

The rest of this document is organized as follows. In Section 2, we provide further historical background, especially regarding medieval Avignon. Section 3 describes how we built a geographic database by combining the information coming from five primary historical sources. We then turn to graphs, by highlighting the interest of such modeling approach in Section 4, and describing in detail the proposed graph extraction methods in Section 5. We compare experimentally these methods by applying them to our Avignon database in Section 6, a process that results in the identification of the optimal confront network. We detect the communities of this graph, which we discuss from a historical perspective in Section 7. Finally, Section 8 summarizes our main contributions and findings, and identifies our most promising research perspectives.

2 Historical Background

After the division of Provence in 1125 [25], the city of Avignon was subject to a complex seigniorial fragmentation. Different powers manifested themselves in the city and many of them gradually acquired rights over the land and urban property. Before the middle of the 14th century, these rights were not clearly defined by all the land lords. Only the greatest of them, and more particularly the Counts of Provence, had surveys carried out to identify the rights and property on their land. It was with the arrival of the papacy in Avignon that all the land lords, whatever their importance in the city, expressed the need to precisely delimit the rights they possessed over the land. The phenomenon became even more pronounced when Pope Clement VI bought the city from the Countess of Provence, Queen Jeanne, in 1348. The local lords had to protect and guarantee their rights, as they were an important source of economic income. It was therefore essential to be able to clearly identify the plots of land that felt under their lordship.

Most of these plots were granted to third parties, called *tenures*. The landowner could indeed transfer the beneficial domain of the property while retaining the eminent domain. The person benefiting from the beneficial domain, whom we will call the *tenant*, could then sell, transfer, or lease the property. The landowner received an annual fee, called *cens*, from the tenant, as well as *transfer fees* in the event of the sale of the beneficial domain.

From the second half of the 14th century on, inventories of tenures multiplied, and census practices were perfected. An increasingly complex listing of the rights of the landed lords, cataloging all the leased tenures under their lordship, was put in place. It took the form of various land registers, the most successful of which, particularly from a legal standpoint, is known today as *terrier* [16].

In the rest of this section, we describe the main characteristics of these land registers (Section 2.1), and the challenges related to their exploitation (Section 2.2).

2.1 Notion of *Terrier*

Starting from the 14th century, most landed lords commissioned the creation of land books. These were real tools intended to manage, authenticate, and protect the rights of landed lords on their properties. They list, in a serial and often stereotyped manner, the properties of those who had to pay a fee to a lord, giving a certain amount of information on the tenant, the land and the amount of the fees to be paid. Each land book consists of a series of declarations, presented as in Figure 1 (left). Each declaration is constructed in the following way. The identity of the tenant is given first (shown in red in Figure 1), and can be defined in different ways. In addition to the person's first and last names, it may contain, for example, their occupation, social status and may be complemented by their family ties, familiarity (i.e. entourage) or professional ties.

Next, the property held by the taxpayer is mentioned (in orange in the figure). In most cases, only the type of property is indicated (house, yard, garden, vineyard, undeveloped land). A more precise description of the plot is sometimes given (state of conservation of the property, building materials, size). The location of the property is then regularly indicated (in blue in the figure). From the 14th century onwards, this location is increasingly precise, while always remaining *relative*. The properties are first of all located in relation to the parish territory; then the street and the *confronts*, i.e. the neighborhood, are mentioned. The due tax is generally indicated at the end of the entry (in green in the figure). A number of different currencies can be used, and the census can even be specified in kind. This makes it difficult to estimate its exact value and to compare taxes from one entry to the other.

Whatever the internal organization of the declarations in the documents, it most often testifies to the increased desire of the institutions to construct genuine management tools intended to be regularly used as legal proof of their possession. These documents are also intended to be archived. In addition to the legal function, most of these documents have a memorial function, which is illustrated above all by their appearance. Some of them are indeed large and beautiful registers, with an elaborate appearance, as shown in Figure 1 (right).

In a city that had become the seat of Christendom, where the Pope had been the true political lord of the city since 1348, and in which numerous land lords held rights over the land, it was now essential that everyone be able to legally assert their possessions. These registers are therefore as much legal evidence as they are demonstrations of the power of the landed lords in relation to the other powers present in the city.

It is impossible to provide an exhaustive list of landed lords who hold rights over urban land in Avignon. While some lords are well known and their rights can be partly identified through produced and preserved documentation, others remain completely unknown to us. Among the most well-known and important landed lords are ecclesiastical institutions, which have multiplied in the urban space of Avignon even before the arrival of the papacy in the city. Moreover, they have shown an early willingness to archive their documentary production. Thus, their documentation –especially that which is the focus of our study, namely the documentation produced by the

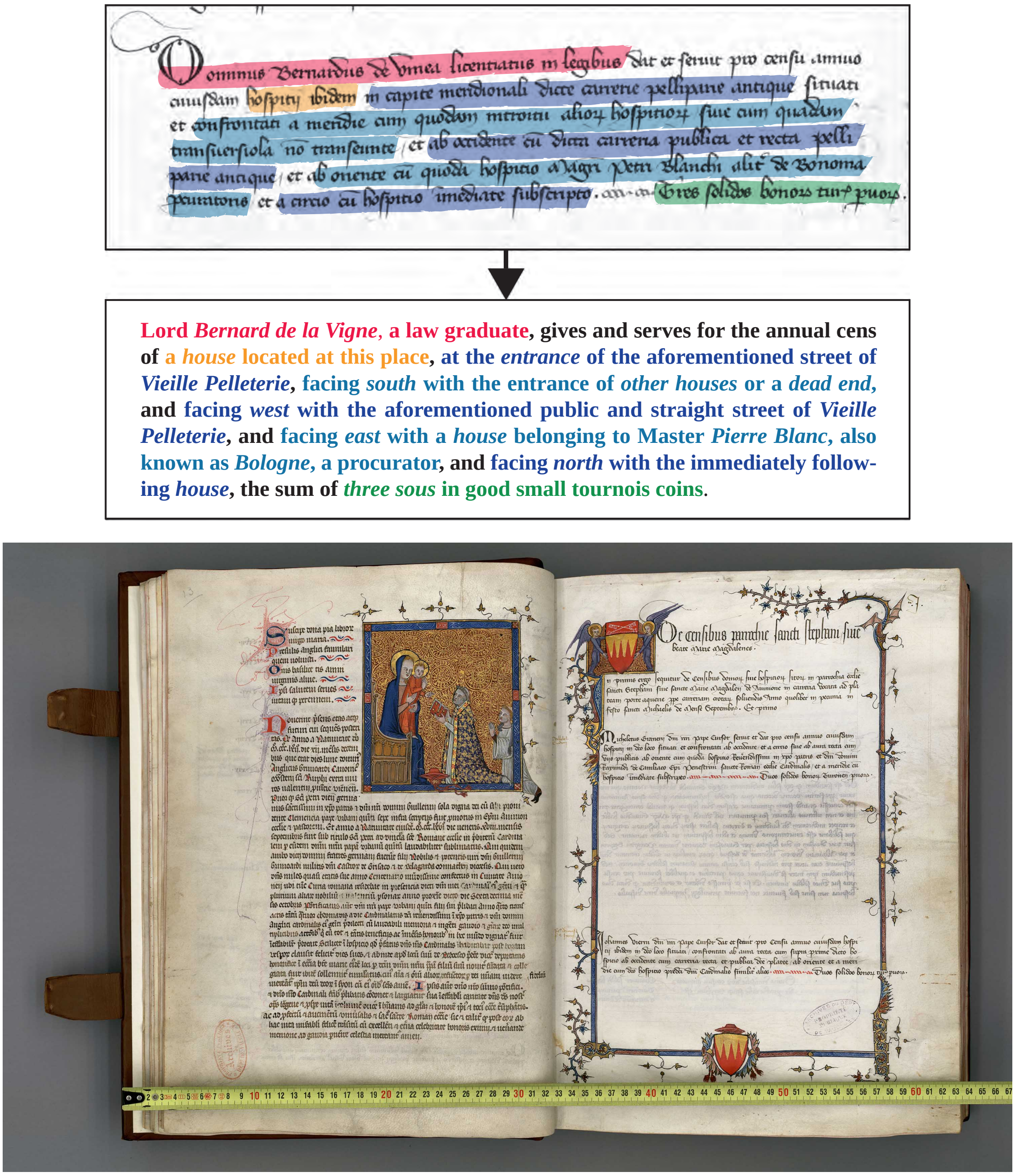

Figure 1: **Top:** Example of declaration retrieved from a terrier, Vaucluse Departmental Archives, 1G10 f.9v. It includes the original text (first frame) and its English translation (second frame). Each color represents a different piece of information: tenant (red), property (orange), location (blue, 5 different confronts here), and fees (green). Italics denote entities of interest. Diagram available at 10.5281/zenodo.14175830 under CC-BY license. **Bottom:** Terrier of Bishop Anglic Grimoard, Vaucluse Departmental Archives, 1G10 f.1.

bishopric, the cathedral chapter, the hospital of Saint-Jean de Jérusalem, the chapter of the Saint Pierre collegiate church, the Sainte-Catherine monastery, and the House of Repentants– is still the best preserved and most accessible in the archive holdings today. Alongside these institutions are illustrious Avignonnais who have acquired significant land lordships through political and economic opportunities. This is the case, for example, of Jean Cabassole, knight and advisor to the Count of Provence Robert d'Anjou. The Count granted him the lordship of a strip of land parallel to the city walls in 1319. Like ecclesiastical land documentation, that produced by the urban community is of considerable richness, and testifies to the manifest desire to preserve the traces of its history and properties. Finally, when the Pope bought the city from the Countess of Provence Jeanne d'Anjou, he became the only political lord of Avignon. On another scale, that of land ownership, he was nonetheless one landed lord among others. The Pope did not have a land registry produced, but we know all of his properties and rights thanks to the accounts of the city's clavigers (a type of municipal officer). These accounts list, in the same way as land registries, all the leased properties belonging to the papal lordship and for which individuals paid a fee.

2.2 Limitations and Difficulties

Terriers are true multiple lists, whose content offers numerous possibilities of analysis. Historians and geographers, particularly those from the *Annales School*, quickly became interested in studying them [5]. However, they faced several obstacles in doing so: 1) quantity and incompleteness of the data; 2) lack of data standardization; and 3) geographic uncertainty.

Data Incompleteness The medieval city is a fragmented territory governed by various intertwined and often opposing powers. The documentation reflects this reality: it is often incomplete and fragmented, rarely covering the entirety of a territory. Studying a city based on these sources therefore requires cross-referencing and comparing the entirety of this particularly extensive documentation. However, the volume of data to process makes this work particularly tedious, and even challenging to approach using traditional historical methods. This extensive amount of data, previously studied in isolation, can now be approached and examined with a fresh perspective thanks to computer power and computational sciences. Moreover, researchers today have access to a substantial number of edited and sometimes digitized written sources, which they can approach in a serial manner.

In Avignon, the first cadastre covering the entire urban area was only created in the late 16$^{\text{th}}$ century. Before this date, it was necessary to gather land books initiated by different lords, each covering different parts of the territory, to obtain the most comprehensive view possible of the urban space. Some areas of the city are better documented than others. Certain zones are not covered by the documentation, either because it has been lost or because it never existed. Some plots of land are indeed free of rights and are fully owned by their owners. In such cases, the holders do not have to pay any fees to a land lord, and these properties are therefore not listed in the land books mentioned above. In this sense, the systematic study of land documentation can also highlight free zones within the city. However, studying these sources systematically requires significant data preparation.

Data Heterogeneity Although terriers share many similarities, the data they contain are not standardized. First, the documents are not all written in the same language. In the case of Avignon, for instance, some are written in Latin while others are in Provençal (the local language at the time). Therefore, the terminology used to refer to the same entity can be very different from one document to another. In the Sainte-Catherine registry, written in Provençal, we encounter Juan Teyseyre; but the same individual is named Johannes Textoris in the Anglic Grimoard registry written in Latin. Moreover, in the medieval period, spelling was not standardized. Significant variations exist even between documents written in the same language. While it is written as Juan Teyseyre in the Sainte-Catherine registry, in the city registry, also written in Provençal, we find the spelling Johan Teysseyre. Thus, considerable work is needed to extract the information from the sources and formalize the data for serial analysis.

Geographic Uncertainty One crucial point that complicates the analysis of this documentation lies in the geographic dimension of the information. Historians have long sought to reconstruct the land division, considering it an essential step in analyzing terriers. This is a very long and tedious task that can only be carried out for limited and particularly well-documented areas [8, 10]. If we want to consider the entire urban space, as we have seen, it is necessary to take into account a vast amount of data and cross-reference them: this work cannot be done manually. Moreover, due to the available information, reliably accomplishing this reconstruction is practically impossible. Indeed, property addressing, street name signage on panels, and street numbering only began to be standardized in the 18$^{\text{th}}$ century, particularly following the Napoleonic conquests. Towards the end of the Middle Ages, informal street names became common; however, they were not standardized and could vary from one person to another or from one document to another. The emergence of street names primarily pertains to the roads that contribute to the political representation of the city and its economic prosperity; the chosen toponyms are a direct testament to this. Most of the names already in use in the 13$^{\text{th}}$ century continued to be used at the end of the medieval period. Nevertheless, there were instances where they were no longer applied to the same street, or where they encompassed multiple streets.

In the land documentation used for this study, the most precise information we have for locating a property, after the street, is that of its boundaries or *confronts*. However, this information does not always allow us to determine the exact position of the property. It provides a *relative* position. It was probably very clear to its contemporaries and left little doubt about the location of the declared parcels. However, today, out of its original context, it can sometimes be completely elusive to us. In fact, when the position of a property is given by its *confronts* with properties owned by various individuals and not with a building, a street, or any other fixed point whose location is known to us, it can be extremely challenging, or even impossible, to precisely locate the properties mentioned in the sources.

3 Geographic Database

The information contained in terriers can hardly be exploited directly, due to their structural and textual heterogeneity, and to their complexity. In the context of this project, we design and apply natural language processing (NLP) tools to extract the relevant information and store it in a proper geographic database. The detail of this process, as well as the exact architecture of the resulting database, are out of the scope of this article: they are provided in [13].

In this section, we first briefly summarize the process proposed to handle this task (Section 3.1), then we describe the obtained spatial objects (Section 3.2) and the spatial relations between them (Section 3.3), which populate our database. Finally, we explain how we complement this database with additional secondary sources (Section 3.4).

3.1 Information Extraction

The comprehensive list of the historical sources used to constitute our database is available in Appendix A. A thorough initial reading of these land registers reveals a certain level of regularity in their structure and in the form of the information they contain. This observation, together with the amount of text to process, the substantial number of edited or already transcribed texts, and the unavailability of any appropriate tool able to handle Latin and Provençal, justifies the elaboration of a custom semi-automatic process aiming at identifying information of interest in the text, extracting it, and storing it in a relational database. Fully automating this process appears unnecessarily difficult, which is why we adopt a method consisting in automating as much of it as conveniently possible, while still relying on human verification and correction in order to control the quality of the data produced in the end.

The first step consists in scanning the edited or transcribed terriers, as no digital editions exist, before performing Optical Character Recognition (OCR) in order to get electronic versions of our historical sources. A manual intervention is necessary to solve the errors introduced by the OCR tool. Note that handwriting recognition tools were not yet sufficiently advanced at the beginning of our study, so we focused on edited or, at a minimum, transcribed documentation that was already substantial. However, handwriting recognition processes are becoming increasingly effective and should soon enable the consideration of even larger documentation.

At this stage, we can apply our NLP tool, called *Auto-Annot* [13]. It relies on a fine-grained categorical model defined based on the studied sources. It is built using a dual symbolic approach, as it both takes advantage of a set of predefined rules describing recurrent patterns and external/internal clues, and of a manually constituted lexicon. The tool is designed to recognize multilingual documents, distinguishing between Latin and Provençal. Entity-wise, it detects the information previously described in Section 2.1 (cf. Figure 1): individuals, fees, properties and other spatial objects. Some entities are described according to several traits, for instance individuals are likely to be described using a first name, family name, nickname, honorifics, hometown, health or legal status. Relation-wise, our tool detects inter-individual links (e.g. family or professional relations) as well as spatial relationships that allow the relative positioning of spatial objects (e.g. being North of a church). The different types of entities and relations targeted in the text were identified during our first reading of the historical sources, and were later complemented iteratively until reaching a satisfying quality level.

The lexical variability characteristic of our historical sources is a major difficulty to detect entities and their relations. The form of Latin used in the late Middle Ages tends to use fewer declensions, but they are still present. Furthermore, the language is marked by processes of vernacularization, including a greater flexibility in sentence construction. Finally, there is no orthographic standardization; quite the opposite. The liberties taken by the writers in choosing vocabulary and spelling are particularly significant. Not to mention the spelling of proper nouns, which sometimes offers surprising variations. It is not uncommon to encounter several different spellings for the same word, or different words used to refer to the same entity within a document. To solve these issues, our tool includes a custom lemmatizer, and a post-processing step leverages Levenshtein's distance to detect similar strings, and more specifically homonyms.

The architecture of our database directly depends on the types of entities and relationships extracted from the text. In addition, it allows storing some information that is not present in the terriers, but rather comes from some secondary sources, such as the absolute location of certain objects (under the form of GPS coordinates). This database allows us to represent geographic information when we have sufficient indications about the location of a historical object and to query the extracted data from the sources more precisely. We next list and discuss the different types of entities and spatial relations between them, and how we exploit the secondary sources.

3.2 Spatial Objects

At the end of the extraction process, our database contains nine types of spatial objects, as listed in Table 1 and detailed in the following.

Table 1: Types of spatial objects present in our database, with their main characteristics. Column *Nbr.* shows their number of occurrences in the database, before graph extraction. Column *Inv.* indicates whether the objects are invariants (cf. text).

Type	**Nbr.**	**Description**	**Inv.**	**Dimension**
Properties	3,021	Privately owned pieces of real-estate (declared and undeclared)	No	Punctual
Parishes / sectors	19	Parochial territories and geographical sectors	No	Surface
Boroughs	59	Extensions of urbanization outside the old city walls	Yes	Surface
Defensive system	7	Moats, ramparts, walls and towers	Yes	Linear
Gates	32	All forms of openings piercing the city walls	Yes	Punctual
Liveries	25	Residences granted to the cardinals	Yes	Surface
Geological landmarks	6	Points of reference such as rocks, rivers, and canals	Yes	Linear/Surface
Streets	326	All types of roads and ways	Yes	Punctual/Linear
Edifices	152	Institutional properties	Yes	Punctual

The first type corresponds to private *properties*, as their description is the main focus of the terriers. It includes houses, backyards, shops, gardens, vineyards, fields, barns, cellars, and others. Figure 2 (left) shows their distribution over the city, based on the historical sources. Some of these properties appear as proper entries in a terrier, as they are explicitly *declared* by their tenants. But some properties are *undeclared*, in the sense that they appear in the terriers only as confronts, i.e. as spatial points of reference used to locate declared properties in a relative way. They might be properly described in other, unavailable documents, though.

The second type of spatial object is *Parishes and sectors*. In the late Middle Ages, Avignon was divided into seven parochial territories of varying sizes. The exact boundaries of these parochial territories are not known to us, as no document references them. However, research conducted by Pierre Pansier [29] and Anne-Marie Hayez [20] has allowed for a proposed reconstruction of these boundaries, as shown in Figure 2 (right). Their cartography reveals significant differences in territorial extent from one parish to another. With urban expansion and the construction of new ramparts in the second half of the 14$^{\text{th}}$ century, some parochial territories experienced considerable enlargement (such as the parishes of Saint-Pierre and Saint-Agricol), while for others, expansion

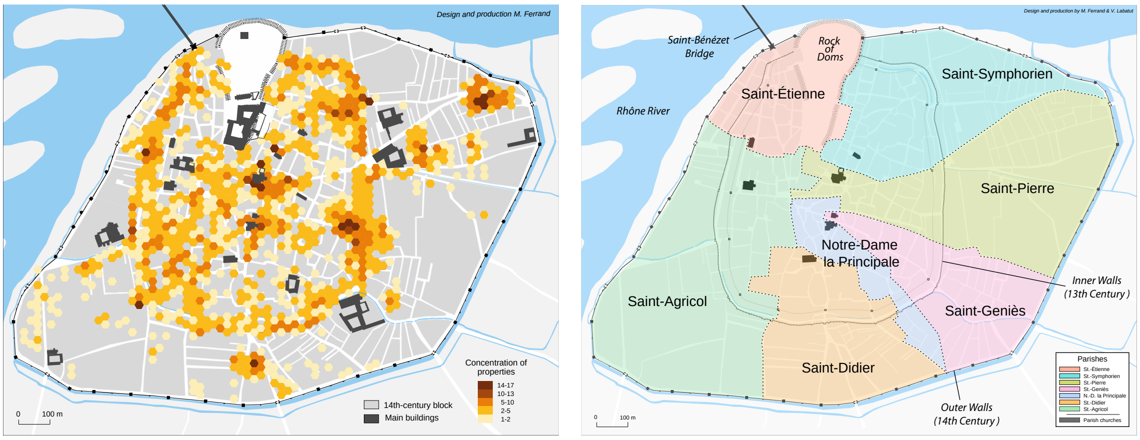

Figure 2: **Left:** Density map of properties (declared and undeclared) in our dataset; location by interpolation using the grid method. **Right:** seven parishes of Avignon, and main geological landmarks. Plots available at 10.5281/zenodo.14175830 under CC-BY license.

was notably limited by the natural landscape (such as the parish of Saint-Étienne). To locate the properties in most of the terriers, the scribes use parish affiliation. However, in some sources, other sectors are used, such as the Jewish quarter, or the space between two gates of the communal ramparts.

The seven remaining object types collectively constitute what we call *invariants*, in the sense that they can be considered as the constants of the urban landscape [32]. This expression refers to all objects that can be referenced in confronts, excluding properties (declared or undeclared). This includes streets, buildings, geological landmarks such as rocks or canals, city gates, and others similar objects.

Boroughs constitute our third object type. By definition, boroughs are "extensions of urbanization outside the city walls." In her research, Anne-Marie Hayez identifies over seventy so-called "bourgs" in Avignon at the end of the 14$^{\text{th}}$ century and locates them approximately. These are subdivisions, meaning that the initially often agricultural lands were later divided to create multiple plots (in terms of housing units). Their size can vary from a few plots to several hundred. Initially, these subdivisions primarily developed outside the walls of the 13$^{\text{th}}$ century. However, in the second half of the 14$^{\text{th}}$ century, they became incorporated into the urban space with the construction of the new city walls.

The fourth object type is the city *defensive system*. The communal ramparts consist of double walls, an inner wall and an outer wall, likely constructed in the 13$^{\text{th}}$ century. Between the two walls are the *intermural* spaces. A moat is also present in front of the outer wall. In the mid-14$^{\text{th}}$ century, new ramparts were built at the initiative of the papacy and the urban community. Each rampart is divided into multiple sections, and has towers. The city *gates* constitute the fifth type of spatial objects. The outer wall of the double city ramparts is pierced by twelve gates. Eight of them have corresponding gates on the inner walls, and in those cases, the facing gates bear the same name. The 14$^{\text{th}}$-century ramparts are also pierced by twelve gates.

The sixth type corresponds to Cardinal's *liveries*. Since 1316 and the regulation of housing, every cardinal who arrives in Avignon is granted a *livery*. These are rented residences that are "delivered," hence the name given to them. The space of a Cardinal's livery's is not limited to a single dwelling that accommodates a cardinal and a few of his entourage. It can be considerable and always contains multiple houses. The number of dwellings that a Cardinal's livery's represents depends on the size and needs of a cardinal's court. To delineate the space assigned to them, the cardinals place barriers, called *cancels*, at the end of the streets that make up their livery. These wooden barriers explicitly enclose the territory of the cardinal's residence.

The seventh type gathers *Geological landmarks*, which correspond to geological points of reference, including rivers, channels, and rocks. The eighth type is *streets*, and it covers not only proper streets, but also alleyways, squares, dead-ends, pathways, etc. Finally, the ninth type is *edifices*. It gathers institutional properties, including administrative buildings, pyres, chapels, graveyards, convents, churches, hospitals, granaries, hospices, monasteries, bridges, wells, etc.

We also distinguish our spatial objects in terms of their *dimensions*. First, some are *punctual* (or 0-dimensional) in the sense that they occupy a relatively compact surface. Most of these are single buildings, such as houses, shops, churches, but also short streets. Second, some objects are *linear* (or 1-dimensional), meaning they cannot be reduced to a single position in space, but rather to a sequence of contiguous positions. These are mainly longer streets, as well as walls, rivers and channels. Third, some objects are better modeled as shapes possessing a surface (2-dimensional), because they are very large and/or are constituted of several adjacent buildings: boroughs, parishes, cardinal's liveries, and certain landmarks. Table 1 shows the dimensionality of each type of object.

3.3 Spatial Relations

In the historical sources, we identify no fewer than 42 distinct types of spatial relationships between these objects. Based on their semantics, one can gather them in six main categories:

- *Cardinal* relationships: being to the north, south, east or west of another object.
- *Vertical* relationships: being above or below another object (e.g. an apartment topping a shop, in the same building).
- *Horizontal* relationships: being behind an object, facing an object.
- *Street* relationships: being at the beginning or end of a street, at the intersection of two streets.
- *Proximity* relationships: being around or near an object, adjacent to an object.
- *Hierarchical* relationships: being inside an object (e.g. a building in a borough), surrounded by an object, just outside an object.

The *Hierarchical* category corresponds approximately to one third of the relations present in the database. It typically indicates that some punctual object is part (or not) of some 2-dimensional object. By comparison, the remaining two thirds are *flat* relationships, in the sense that they locate an object relative to another object without any notion of inclusion (e.g. *West of*). A comprehensive list of all 42 spatial relation types is provided in Appendix B (Table 6).

3.4 Additional Information

If all the objects described in the land registers, along with their relationships, are inventoried in the database, it is not possible to determine their absolute spatial location, due to insufficient information. Drawing on numerous studies focusing on the urban space of Avignon and using a regressive approach to the landscape based on available planimetric sources and archaeological data, the majority of the invariants, however, have been located in the literature. A proposal for the restitution of the parish territory [20] and the medieval road network has been made [29], areas of boroughs [19] and cardinalatial liveries [21] have been identified, and the extent and geographical position of the main buildings have been highlighted. We leveraged these results, and then refined and supplemented them.

We have georeferenced a large part of our spatial objects in a Geographic Information System: 50 out of 59 boroughs, all parishes and geographical sectors, the entire defensive system and city gates from different eras, all geological landmarks, all Cardinal's liveries, 162 our of 326 streets, and 144 out of 152 edifices. It should be noted that while we have georeferenced these objects, only a small portion of them are accurately located. The level of accuracy depends on whether they are still in place, which is the case of certain churches and buildings, and on whether textual, iconographic, and especially archaeological sources allow us to place them precisely. For the others, we rely primarily on a hypothetical location made possible by the compilation of various sources and historical studies.

We matched the declared properties identified in our corpus with the finest spatial reference for which we had information (parish, borough, street), and then arranged them manually based on additional information, particularly confronts. Out of 3,021 properties, 2,049 have been georeferenced. However, this property localization is highly uncertain because it relies on the localization of all the invariants, which is often quite hypothetical. Therefore, it is nearly impossible to locate the properties without errors. It should be noted that we do not have addresses but only relative positions. Attempting to precisely locate the properties inevitably leads to data degradation.

As a result, we decided to move away from georeferencing the properties and focus on the main historical source of information, which is the topological relationship between the properties. Building on existing historical research and various sources, including planimetric sources used to spatialize certain objects, we were able to enrich our database and add certain relationships between objects that are not present in the land documentation. Data related to the road network, in particular, were enhanced with information about the connections between each street or square. Finally, we also added information about the positioning of the buildings, specifying the street in which each listed building is located. To refine the information, the streets were then divided into segments, and each of these segments was linked in the database. Similarly, the Rock of the Doms, a very important and frequently used landmark, was broken down into several pieces that were interconnected in the database. This allowed for a more detailed topological refinement of the original information.

In the rest of the document, we refer to the information extracted from the land registries as *primary* data, whereas that produced from secondary sources through the process described above is called *additional* data.

4 Graph-Based Modeling

To address the issues of data incompleteness and geographic uncertainty that affect terriers, we leverage graph theory to model the urban space. In this section, we justify this methodological approach (Section 4.1) before discussing the challenges it implies (Section 4.2).

4.1 Advantages of a Graph Model

Graphs are designed to represent relational information, therefore they are particularly well-suited in our case, where the retrieved data are essentially a collection of spatial objects with spatial relations between them. These objects can be modeled as vertices, and these relationships as edges between them. Moreover, graphs enable us to overcome the need for precise property locations. If the absolute position of a spatial object could be estimated, then this information can be included as coordinates attached to the corresponding vertex, making the graph *spatial*. Otherwise, this vertex simply does not have any known position. In addition, graphs even allow us to handle certain contradictions present in the historical sources, e.g. some building being located to the North of another in some terrier entry, whereas the latter is also declared to the North of the former in another entry. When one does not know which piece of information is correct, keeping both seems to be a reasonable approach. Finally, another important advantage of graphs, compared to the more traditional modeling approach, is that they exempt us from the tedious (or even impossible) task of reconstructing parcel divisions, mentioned in Section 2.2.

Not only does graph-based modeling not require reconstructing parcel divisions, but it even allows partitioning the urban space at a higher-level. In the Middle Ages, one of Avignon's characteristics is the absence of any administrative subdivisions other than the parish. The parish is indeed the religious subdivision of the city, but it is also the only civil division used simultaneously from an administrative, fiscal, and military perspective. However, it is a fairly large territory, directly affected by urban expansion, and ultimately not very representative of socio-spatial relationships. Two individuals can belong to the same parish without necessarily living in proximity. Two people can be attached to the same parish without any actual connections. To study the socio-spatial relationships of proximity and the resulting land uses, we must therefore use another scale that is closer to the individual himself. A graph extracted from our database defines a certain representation of the urban space, very close to the declarers' perception, as it is based on spatial relationships declared by the tenants themselves. In this regard, we will see that partitioning such a graph by the means of a community detection method [15] allows for a more refined scale of analysis. In this case, communities correspond to *ad hoc* neighborhoods, built upon the confront relations. Here, we use the word *neighborhood* in the sense of a lived space, that is, as it is perceived and practiced by individuals [24]; its definition is based on the topological relation defined in the sources. The neighborhood is therefore primarily understood as the everyday living space in which relationships between individuals are the densest.

4.2 Graph Extraction Challenges

As we will see in Section 5, due to the richness of the data collected in our database, there are a number of possible ways to extract graphs representing medieval Avignon. They differ mainly in the data that they leverage: which part of the primary historical data to keep or remove, and which additional secondary information to use. In order to make a choice, it is important to define a method to compare them and identify the most appropriate for our use.

The first goal is to have a graph containing as many vertices as possible, so that it covers the largest possible area. Put technically, we want to maximize the graph order. More precisely, we want to maximize the number of properties it contains, as these spatial objects constitute our main interest in this study. At the same time, we want this graph to be as dense as possible, so that its structure reliably reflects the spatial organization of the city. Yet, as was observed empirically, complex networks tend to be sparser as they get larger [4], which means that these two goals of coverage and reliability may be opposed.

Finally, it is worth stressing that our database describes a number of very different spatial relationships, with very different levels of reliability. Therefore, the extracted edges may not all represent the spatial structure of the city equally, in terms of semantics, granularity, precision, etc. Thus, another challenge is to determine how to build a graph structure that constitutes a good approximation of the spatial structure, in order to produce a relevant partition of the city as a final result.

5 Graph Extraction Methods

The most straightforward method to extract graphs from our database is to leverage the whole information, and model each object by a vertex, and each relationship by an edge. This is illustrated by a simple example in Figure 3. The left part of the *Database* block represents the spatial relations described in the terrier, whereas its right part shows the absolute locations of the listed objects, when available. Each color represents a unique spatial object, some of which are properties. In the *Terrier* part, the left-hand squares are properties, associated (on their right) to their list of declared confronts. Some objects appear only because they are cited as confronts, e.g. the yellow one. The absolute spatial location of certain objects is unknown, e.g. the green property does not appear in the right part of the *Database* block. Graph edges do not necessarily depend on spatial proximity, since they reflect the confronts, i.e. the relationships that are *explicitly* described in the database (left part). For instance, the magenta and yellow objects are spatially close, but not connected by any confront, so there is no edge between the corresponding vertices.

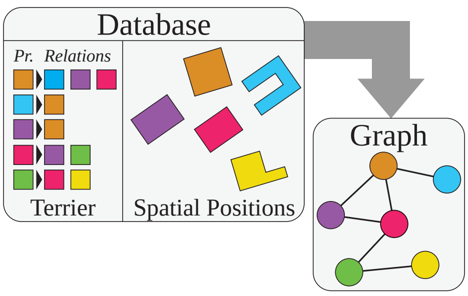

Figure 3: Straightforward extraction of a graph, from our database. Each colored shape represents a spatial object (e.g. a building). The edges depend on the spatial relations described in the historical sources. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

Our database describes various types of objects (e.g. buildings, cemeteries, walls, gardens, parishes), spatially related in various ways (e.g. being near another object, containing another object). Leveraging the entirety of this information results in a very heterogeneous graph, as it is simultaneously spatial [2], attributed [34], multimode [7], and multilayer [23]. It is shown in Appendix C (Figure 11), for reference. However, not all of this information may be useful to produce a graph that would provide a good approximation of the spatial distance. In fact, the preliminary experiments that we conducted revealed that some of this information can even be detrimental. Consequently, one of our methodological challenges consists in identifying which part of the data should be discarded. On the contrary, it is possible to leverage some additional information besides that strictly coming from the terriers, by using our secondary historical sources, and possibly improve the quality of the extracted graphs. We implement these different approaches under the form of several distinct extraction steps. These are not mutually exclusive, they can be chained in various ways, resulting in a number of possible combinations. We consider some of these steps to be *compulsory*, because the extracted graph is hardly usable at all without them: we describe them in Section 5.1. Some are *optional*, and we later study how they affect the extracted graph: these are the object of Section 5.2. Finally, we provide the reader with a convenient summary of these methods in Section 5.3.

5.1 Systematic Steps

Both steps described below are systematically applied to all extracted graphs. They concern the way we normalize relationships, and how we handle disconnected graphs.

5.1.1 Relationship Types

The number of distinct relationship types (42) is an issue, as they complicate the interpretation and analysis of the network. Moreover, some of them are very infrequent, occurring only once or twice out of the 6,129 relationships listed in the database. For this reason, we normalize these types to make them coarser, retaining only 7 main types: being to the *North*/*South*/*East*/*West of* an object (i.e. so-called cardinal relationships), being *Inside*/*Outside* an object, and being *Around* an object (a general relationship gathering all the remaining types). The detail of this normalization is provided in Appendix B (Table 6).

5.1.2 Minor Components

Generally speaking, real-world complex networks often contain a so-called *giant component*, i.e. a very large maximally connected subgraph gathering almost all the vertices and edges, whereas the rest are spread over several much smaller separate subgraphs [26]. It is also the case for all the networks extracted from our database, whatever the extraction method. In this situation, the standard approach adopted in the literature is to focus the analysis only on the giant component, or on the few largest components, as the rest of the data are implicitly considered as less complete or reliable. In our case, more particularly, approximating the spatial distance on these remote and sparser parts of the network is likely to result in a poor estimation. For this reason, for each extracted network, we filter out all vertices and edges located in minor components, including isolates. For our data, we empirically determine that a lower threshold of 25 vertices is appropriate to remove noisy data while preserving relevant information.

5.2 Optional Steps

We perform our extraction procedure by alternatively including and discarding the following steps, in order to assess how they affect the extracted graphs. While describing these steps, we also introduce the notation used later to refer to them and their different variants.

5.2.1 Non-Punctual Objects

The vertices representing 1- and 2-dimensional objects are likely to act as shortcuts in the network. For instance, two buildings located at opposite ends of a street are just two hops away in terms of graph distance, independently of the street length. The blue and pink buildings in Figure 4 illustrate this situation (*Keep* graph). This could strongly affect the way our confront graph models the urban space. We propose two modifications of the extraction process to deal with this issue: outright removing the concerned vertices (*Remove* graph in Figure 4), or splitting them (*Split* graph).

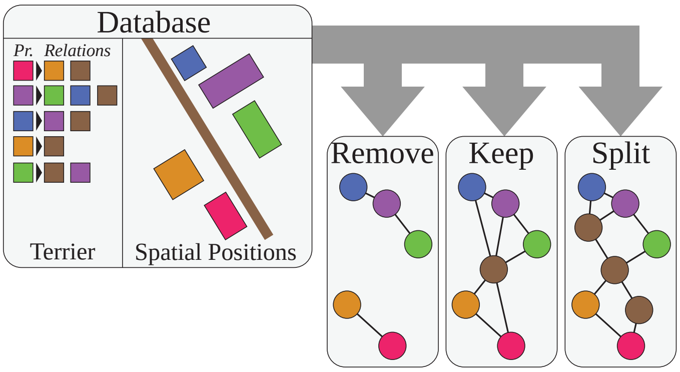

Figure 4: The three strategies proposed to handle 1- and 2-dimensional objects such as the street in this figure (shown in brown): removing it by not representing it at all in the graph, keeping it as it is by modeling it through a single vertex, or splitting it and representing its constituting pieces with several linearly connected vertices. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

Removal Removing *all* non-punctual objects is the most straightforward approach. However, our experiments reveal that this causes the networks to be very segmented, i.e. it breaks them down into many small components. Consequently, they may not be suitable to produce acceptable approximations of the urban space.

A less radical approach consists in removing only the vertices representing the longest/largest objects. Of course, this requires possessing the required spatial information, which is not always the case. Our database contains the length of certain streets, and we are confident that this information is missing only for the most minor streets, which are also the shortest. In addition, a few relationships also specifically concern parts of streets (angle, beginning, end, etc.).

We experiment with two removal strategies[1]: 1) keep all streets (a strategy we denote `streets`); 2) remove the k longest streets, and keep the rest of the streets (noted `k`). In both strategies, we outright remove the other types of 1- and 2-dimensional objects.

Split Alternatively, we consider splitting non-punctual objects instead of removing them, i.e. breaking them down into smaller parts, each one represented by a specific vertex. This requires injecting more knowledge in the database than the previous method, though. First, if the spatial location of the original object is known, it is necessary to estimate those of its constituting pieces. Second, to preserve continuity, we must introduce artificial relationships between the adjacent pieces of the split object[2]. Third and finally, relationships that involve split objects must be adjusted by identifying which piece of the original object they concern.

In our case, we are able to manually split all linear objects (streets, rivers, channels, walls) in segments sufficiently small to be considered as punctual. We also break down the largest 2-dimensional object (the so-called *Rock of the Doms*). We experiment with two splitting strategies: 1) split all streets (noted `streets`); and 2) split the k longest streets, and keep the rest of the streets as single vertices (noted `k`). Like before, the rest of the non-punctual objects are removed.

We use the letters `W` (*whole*) and `S` (*split*) to denote whether or not a graph underwent such a splitting step during extraction.

[1]We actually considered a third strategy consisting in removing only long streets and keeping short streets and parts of streets. It is not presented here because the results are quite similar to the `k` method, but these are included in our data repository.

[2]We iteratively remove vertices possessing exactly one such artificial relationship, though, as these leaves make the graph unnecessarily larger without bringing any useful information.

5.2.2 Hierarchical Relationships

As explained in Section 3.3, our database includes hierarchical and flat relationships. The issue here is that 2-dimensional objects, which are involved in the former type of relations, are all very large in our case. Including hierarchical relationships in the extracted graph results in the presence of hubs corresponding to these objects, connected to many punctual objects. Some of these are very far from each other in terms of spatial distance, but are very close in the graph, as one is only two hops away from the other. Consequently, the graph is likely to constitute a poor approximation of the urban space, as already observed when discussing the split of linear objects. This is illustrated in Figure 5, which shows a parish (in gray) containing eight buildings (in colors). Independently of the spatial distance between them, considering the hierarchical relationships leads to a graph where they are all two hops away from each other, as they are all connected to the parish vertex.

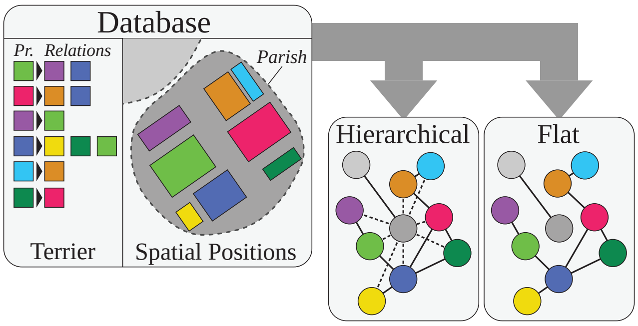

Figure 5: The two strategies proposed to handle hierarchical relationships, such as a parish containing eight buildings in this figure: keep them and make the graph hierarchical, or remove them and make it flat. The dotted edges show the hierarchical relationships. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

We propose to experiment by outright deleting all hierarchical relationships when extracting graphs, a strategy which we note `F` for *flat*, whereas keeping all relationships is noted `H` for *hierarchical*. It is worth stressing that even when focusing on flat relationships only, the membership information can still be retained in a non-structural way, by integrating it in the graph under the form of vertex attributes. For instance, a given vertex can be labeled with the district, parish or borough to which it belongs.

5.2.3 Additional Relationships

As explained in Section 2.1, the terriers list and describe only certain declared properties, and only indirectly mention other objects for the purpose of locating them. They therefore ignore objects that were not recognized during this period —these could be properties belonging to another lord for which no such document is preserved, or properties free from feudal jurisdiction-– or objects that were never mentioned in these documents (such as certain religious and civil buildings, for example). Furthermore, a relationship extracted from a terrier necessarily concerns a property declared in the terrier and some object mentioned in the same terrier. This description is likely incomplete: the authors may not list all possible relationships but only the few necessary to accurately locate the property. Additionally, there is no relationship between two objects that are not declared in these documents, for example, two streets. In summary, our primary historical sources provide an incomplete view of the city.

The question is to know whether this is enough to get a reliable estimation of the spatial distance through the extracted network. In order to answer this question, we leverage the additional information described in Section 3.4, in order to complement our graphs with two types of missing relationships: between adjacent streets, and between edifices and adjacent streets. Figure 6 shows how this affects the graphs. The pink part of the *Database* block corresponds to the additional data used to extract additional edges, shown in red in the *Extended* version of the graph. Our goal here is to study the effect of this extra information on the estimation of the distance. We note `R`

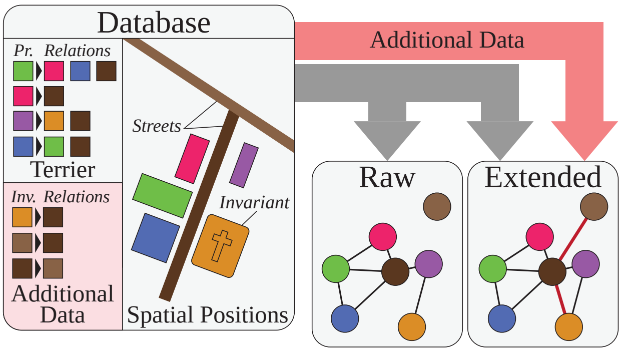

Figure 6: Extending the perimeter of the considered historical sources allows including additional relationships, represented here in red, between streets and other invariants. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

(for *Raw* data) the version of our networks without this information, and `E` (for *Extended* data) the version that includes it.

Table 2: List of the different methods proposed to extract spatial graphs from our database, with the processing steps they rely upon. The letters in the code stand for: use only raw data (`R`) vs. extend them with additional relationships (`E`); use only flat relationships (`F`) vs. include also hierarchical ones (`H`); split large objects (`S`) vs. keep them whole (`W`).

Method Code	**Additional Relationships**	**Hierarchical Relationships**	**Non-Punctual Objects**
`RHW_all`	No	Keep all	Keep all
`RFW_all`	No	Remove all	Keep all remaining non-punctal objects
`RFW_streets`	No	Remove all	Keep only streets, remove other non-punctual objects
`RFW_k`	No	Remove all	Remove k longest streets and other non-punctual objects
`EHW_all`	Yes	Keep all	Keep all
`EFW_all`	Yes	Remove all	Keep all remaining non-punctal objects
`EFW_streets`	Yes	Remove all	Keep only streets, remove other non-punctual objects
`EFW_k`	Yes	Remove all	Remove k longest streets and other non-punctual objects
`RHS_all`	No	Keep all	Split all
`RFS_all`	No	Remove all	Split all remaining non-punctal objects
`RFS_streets`	No	Remove all	Split only streets, remove other non-punctual objects
`RFS_k`	No	Remove all	Split k longest streets, keep other streets, remove the rest
`EHS_all`	Yes	Keep all	Split all
`EFS_all`	Yes	Remove all	Split all remaining non-punctal objects
`EFS_streets`	Yes	Remove all	Split only streets, remove other non-punctual objects
`EFS_k`	Yes	Remove all	Split k longest streets, keep other streets, remove the rest

5.3 Overview

The optional processing steps described in this section are not mutually exclusive: on the contrary, they are meant to be combined. This results in a number of possible methods to extract a network from our databases. The goal of our experimental work is to identify which one is the most in line with our objectives. Table 2 summarizes the combinations of these steps that we consider in our experiments, with the code that we use in the rest of the article to refer to them. We do not consider all possible combinations, because some result in relatively similar effects, even if not exactly identical. For instance, removing all 2-dimensional objects is very similar to removing all hierarchical relationships, as all of the latter involve the former. In the rest of the article, we use a wildcard • to denote any variant over some column(s) of Table 2. For instance, `R•W_all` denotes collectively `RHW_all` and `RFW_all`.

6 Method Comparison and Selection

We apply all 16 variants of the extraction process described in Table 2, and get as many versions of the network representing medieval Avignon. All of them are shown in Appendix C (Figures 12 and 13), as well as the full graph (Figure 11), for reference. The latter is the graph extracted when using the raw dataset, without any vertex or edge removal, or any transformation. Table 3 shows the main topological properties of these graphs. In the following, we first discuss methods to measure the coverage of the extracted graphs, and their ability to reliably model the spatial structure of the data (Section 6.1). We then assess separately the effect of each optional processing step described in Section 5.2: hierarchical relationship filtering (Section 6.2), non-punctual object filtering (Section 6.3), vertex splitting (Section 6.4), and data extension (Section 6.5). Finally, based on these results, we select the most appropriate graph to perform our analysis (Section 6.6).

Table 3: Main topological characteristics of the networks extracted from our database: numbers of vertices (n) and edges (m), density (δ), number of property vertices and proportion of such vertices relative to the database (*Properties*), number of components ($C(G)$), diameter ($d_{\max}$), harmonic mean of the graph distance ($\langle d \rangle$), Spearman's correlation between the graph and spatial distances (ρ_d).

Method	n	m	δ	# Properties		$C(G)$	$d_{\max}$	$\langle d \rangle$	ρ_d
Full graph	3,173	6,619	0.0007	2,693	100,00%	110	16	6.75	0.22
`RHW_all`	2,867	6,415	0.0008	2,397	89.01%	1	16	5.51	0.29
`RFW_all`	2,174	4,290	0.0009	1,807	67.10%	5	45	13.12	0.03
`RFW_streets`	2,003	3,980	0.0010	1,673	62.12%	12	33	38.22	0.48
`RFW_k`	1,903	3,750	0.0010	1,597	59.30%	14	40	45.10	0.49
`EHW_all`	2,919	6,840	0.0008	2,427	90.12%	1	15	5.46	0.34
`EFW_all`	2,390	4,895	0.0009	1,959	72.74%	1	26	8.28	0.48
`EFW_streets`	2,268	4,647	0.0009	1,862	69.14%	1	30	9.37	0.74
`EFW_k`	2,167	4,382	0.0009	1,782	66.17%	2	31	11.28	0.80
`RHS_all`	3,074	6,630	0.0007	2,397	89.01%	1	21	6.19	0.47
`RFS_all`	2,381	4,505	0.0008	1,807	67.10%	5	75	21.57	0.27
`RFS_streets`	2,032	4,017	0.0010	1,673	62.12%	12	34	40.79	0.49
`RFS_k`	2,020	3,578	0.0009	1,673	62.12%	12	33	40.32	0.49
`EHS_all`	3,146	7,072	0.0007	2,427	90.12%	1	22	6.13	0.48
`EFS_all`	2,617	5,127	0.0007	1,959	72.74%	1	51	11.25	0.74
`EFS_streets`	2,317	4,701	0.0009	1,862	69.14%	2	34	11.86	0.69
`EFS_k`	2,294	4,208	0.0008	1,862	69.14%	1	37	10.32	0.80

6.1 Measuring Coverage and Reliability

The numbers of vertices (n) and edges (m) vary significantly between the different versions of the graph. On the contrary, their densities (δ) are comparable, showing a level of sparsity that is on par with what is often observed in general real-world networks of the same order [4, 12]. The *Properties* column contains two values: the number of vertices representing properties in the considered graph, and the proportion of properties in the whole database that this number amounts to. As such, these statistics are related to our *graph coverage* criterion described in Section 4.2: we want these values to be as large as possible. The number of connected components ($C(G)$) is largely affected by the filtering step, as shown by the marked difference between the full graph (110 components) and the other graphs (at most 14 components). A large number of components reflects a very segmented representation of the city space, so this metric is related to how reliably the graph models this spatial organization.

At first sight, the diameter ($d_{\max}$), which is the largest geodesic distance observed in the graph, could be considered as related to graph coverage: a large value means the graph is "wide". However, it is not the case for two reasons. First, this width is only topological, and not necessarily spatial: removing certain vertices can make a graph more linear, thereby increasing its diameter without covering more space. Second, the geodesic distance is traditionally computed by discarding pairs of vertices located in different components, as they are conventionally considered as infinitely distant. Therefore, removing certain vertices from a network can also result in splitting it in several components, thereby decreasing the diameter without necessarily changing its spatial coverage. Both situations appear in Table 3, which confirms that the diameter is not an appropriate metric in our case. The average geodesic distance ($\langle d \rangle$) is more interesting, provided one computes the *harmonic* mean instead of the traditional arithmetic mean. As noted by Newman [26], it allows taking into account infinite distance values in the computation.

The simplest way to assess how reliably the graph models the spatial structure of the data is probably to measure the statistical association between the spatial distance, i.e. the Euclidean distance between the spatial objects themselves, and the graph distance, i.e. the distance on the graph between the vertices representing these objects. We experiment with Kendall's τ and Spearman's ρ coefficients, both widespread in the literature, because such rank correlation measures allow taking infinite values into account (unlike the more widespread Pearson's coefficient). We get qualitatively similar results with both of them, so in the rest of the article we only present the results obtained with the latter, which is faster to compute. These values are shown in the last column of the table (ρ_d), and it appears that the graphs exhibit a wide range of correlation values. The full graph, which, by construction, has maximal coverage, reaches only a 0.22 score, whereas some of the smaller graphs get 0.80 scores.

In the following, we discuss how each extraction step presented in Section 5.2 affect the coverage and reliability of the graphs.

6.2 Flat vs. Hierarchical Relationships

We first consider whether to keep or remove hierarchical relationships, i.e. edges that model the inclusion of some spatial object into some other object, e.g. a building belonging to a parish. In each of the four parts of Table 3, the first two rows compare these two options (`•H•_all` vs. `•F•_all`). Independently of the other extraction steps, retaining the hierarchical relationships results in a much higher number of properties, i.e. a much better coverage. For the reliability, on the contrary, it depends on whether we stick to the raw data or leverage the additional relationships (`R••_all` vs. `E••_all`): in the former case, the distance correlation, and thus the reliability of the graph, is higher when using hierarchical relationships, whereas in the latter case it is lower. But even when the hierarchical relationships bring a better distance correlation, we must stress that this value is not satisfying, as it is closer to zero than to one.

Explaining these observations is quite straightforward by looking at the other statistics. When using only the raw data, removing the hierarchical relationships segments the graphs in several separated components (see $C(G)$). Many of them are small, and thus discarded at the filtering step, which causes the lower coverage. Moreover, the removal of the hierarchical edges also breaks shortcuts, as shown by the largely increased diameter ($d_{\max}$) and average distance ($\langle d \rangle$). This is quite visible when considering the full graph (Figure 11), as it shows how the pink edges modeling these relationships connect local hubs (parishes and boroughs) to their neighborhoods. The graph representations from Figure 12 use a geographic instead of an algorithmic layout, and shows that these relationships have a very long range, spatially speaking. This means that they connect spatially remote vertices, making them closer in the graph than they should be, as hypothesized in Section 5.2.1. This is the cause for the relatively low distance correlation.

In summary, hierarchical relationships allow a very good coverage, but seriously hinder the reliability of the graph. We conclude that it is preferable to ignore this type of relationships during graph extraction. In the rest of this section, we focus our discussion on the flat graphs (`•F•_•`).

6.3 Keeping vs. Removing Non-Punctual Objects

The next extraction step is the possible removal of non-punctual objects. As a consequence, we focus on the top half of Table 3, which corresponds to whole vertices (`•FW_•`), by opposition to split vertices in the bottom half, because splitting induces a different processing of non-punctual objects. Such objects include a few 2D entities (e.g. parishes), but there are predominantly linear objects (i.e. 1D), and more particularly streets (see Table 1).

We first compare between retaining all such objects (`•FW_all`) vs. removing all non-punctual objects except streets (`•FW_streets`). This removal causes a significant decrease of the graph coverage: this is expected, as deleting parishes and other 2D objects indirectly results in the removal of the hierarchical edges, and therefore yields the same effect as observed in Section 6.2. This effect is even stronger as one removes more vertices and edges. When using only the raw data (`RFW_streets`), the number of components and the average distance increase notably, which is not the case when using the additional relationships (`EFW_streets`). In both cases though, the distance correlation is much higher compared to the graphs obtained when retaining all non-punctual objects (0.03 vs. 0.48, and 0.48 vs. 0.74, respectively). The explanation is the same as before: non-punctual vertices are likely to connect objects that are spatially very distant, thereby decreasing the graph reliability. The Rhône river constitutes an extreme case of this issue, see the top-left vertex in the graphs of Figure 12.

Some of the streets are quite long, and likely to cause the same problem, which is why we experiment with their controlled removal through methods `•FW_streets`. As explained in Section 5.2.1, they consist in removing the k longest streets (in addition to the other non-punctual objects) while keeping the shorter streets. In order to determine the best value of k, we rank all streets by decreasing length and iteratively remove an increasing number of streets while computing the same metrics as in Table 3. Simultaneously maximizing coverage and distance correlation is a bi-objective optimization problem. The corresponding plots are provided in Appendix D.1 (Figure 14). The most appropriate solutions appear to be the removal of the $k = 6$ and 7 longest streets for `RFW_k` and `EFW_k`, respectively. When comparing the former to `RFW_streets` (i.e. keeping all streets), the difference is hardly noticeable: a small decrease in coverage (-5%) and a small improvement in distance correlation ($+2\%$). However, when comparing `EFW_k` to `EFW_streets`, we see a comparable decrease in coverage (-4%) and a much improved correlation ($+8\%$).

In summary, removing all the non-punctual objects except streets leads to a slightly lower coverage, but a much improved reliability. Removing also the longest streets has not much effect when dealing with the raw data, but improves the results further when using the extended data.

6.4 Whole vs. Split Vertices

Next, we consider the splitting process described in Section 5.2.1: it consists in using several separate vertices to represent portions of 1D and 2D objects, instead of just removing them outright. We focus on the bottom half of Table 3, that shows the results involving this optional step (`•FS_•`). A street is sometimes the only vertex connecting a property to the rest of the graph. Thus, removing a vertex representing a street may indirectly lead to the deletion of properties from the network. For this reason, our expectation is that splitting instead of removing will allow improving the coverage. Moreover, splitting a spatially large object into several vertices increases the length of the concerned paths on the graph, which is likely to improve the distance correlation, too.

We first consider both methods that split all non-punctual objects (`•FS_all`). As expected, when compared to their counterparts that keep all these objects whole (`•FW_all`), it appears that splitting has no noticeable effect on coverage, as the proportions of properties are identical: 67.10% for both `RFW_all` and `RFS_all`, and 72.74% for both `EFW_all` and `EFS_all`. On the contrary, there is a clear increase of the distance correlation: 0.03 vs. 0.27 (`RF•_all`), and 0.48 vs. 0.74 (`EF•_all`). This is because splitting object such as long streets increases the length of certain paths on the graph, as exemplified by the ramparts that appear clearly in Figure 13 (as yellow edges around the city). When we split only the streets and remove the other non-punctual objects (`•FS_streets`), the observations are roughly the same.

We now focus on the last methods (`•FS_k`), which consist in splitting the k longest streets, keeping the shorter ones, and removing the rest of the non-punctual objects. Unlike with the methods based on vertex removal (`•FW_•`), estimating the best value of k is not a bi-objective optimization problem, because splitting an increasing number of streets does not affect the coverage (cf. Figure 15). Consequently, we just select the values that maximize distance correlation: $k = 6$ (`RFS_k`) and $k = 7$ (`EFS_k`). Compared to their counterpart methods that remove the longest streets instead of splitting them, we observe a better coverage while retaining the same distance correlation.

To conclude, our experiments show that splitting non-punctual objects instead of removing them allows preserving the property coverage, and sometimes even improving it, while increasing the distance correlation when it is low, or preserving it when it is already high.

6.5 Raw vs. Extended Data

It is possible to complement the data extracted from the original sources with some additional information (cf. Section 3.4) in order to insert extra relationships in the graphs (cf. Section 5.2.3). However, this operation has a cost. Hence, the interest of assessing whether such effort allows improving the extracted graph. For this purpose, in this section we compare the 1st and 3rd parts of Table 3 (`R••_k`) with their extended counterparts in the 2nd and 4th parts (`E••_k`) of the same table.

The effects are the same independently of the other optional steps, and they are quite strong. First, using the extended dataset considerably improves the coverage: the increase ranges from 5 to 7 percentage points. Second, this operation also greatly improves the reliability, as the increase in distance correlation ranges from 0.26 to 0.45. This improvement also shows in the number of components ($C(G)$), diameter ($d_{\max}$), and average graph distance ($\langle d \rangle$), which systematically decrease.

In conclusion, our experimental results show that making the effort of gathering the additional data describing spatial relationships between objects that are not properties, allows to significantly improve the extracted graph, whatever other optional steps are used during this process.

6.6 Concluding Remarks

Based on the analysis conducted in this section, we can now identify the best graph extraction methods among the proposed variants. Our main observations are as follows:

1. Hierarchical relationships are largely detrimental to network reliability, and should be removed.
2. Additional relationships always improve coverage and reliability, and should be used when available.
3. Removing non-punctual objects strongly improves reliability while only slightly decreasing coverage, provided streets are kept whole.
4. Splitting non-punctual objects instead of removing them leads to much better reliability, for the same coverage.
5. Splitting only the longest streets, keeping the remaining streets whole, and removing the rest of the non-punctual objects, seems to be the best trade-off *when using additional relationships.*

We use Figure 7, which shows how each method performs in terms of coverage (number of properties in the graph) and reliability (distance correlation) as a visual aid to make our final decision. To go beyond distance correlation, Figure 16 (Appendix D.2) shows the relation between graph and spatial distances. We consider that the method exhibiting the best trade-off overall is `EFS_k` (extended relationships, flat relationships only, split the longest streets, keep the shortest, and remove the rest of the non-punctual object), and we will discuss the corresponding graph in the next section. However, this method requires some extra data, both to split the streets and to include additional relationships. In case one is not able to perform the split step (methods shown in blue in the figure), we consider that the next best choice is `EFW_k`. If no additional relationship is available (in green), then graph quality seriously drops, and the most conservative approach seems to be the best choice: `RHS_all`. Finally, if it is not possible to split nor use the extended relationships, then `RFW_streets` seems to be the best method, even if it requires removing many vertices before reaching a passable reliability.

7 Segmentation of the Spatial Graph

At this stage, we have identified the most appropriate method (`EFS_k`, see Section 6.6) to extract a spatial graph representing the city of Avignon in medieval times. We now proceed with the analysis

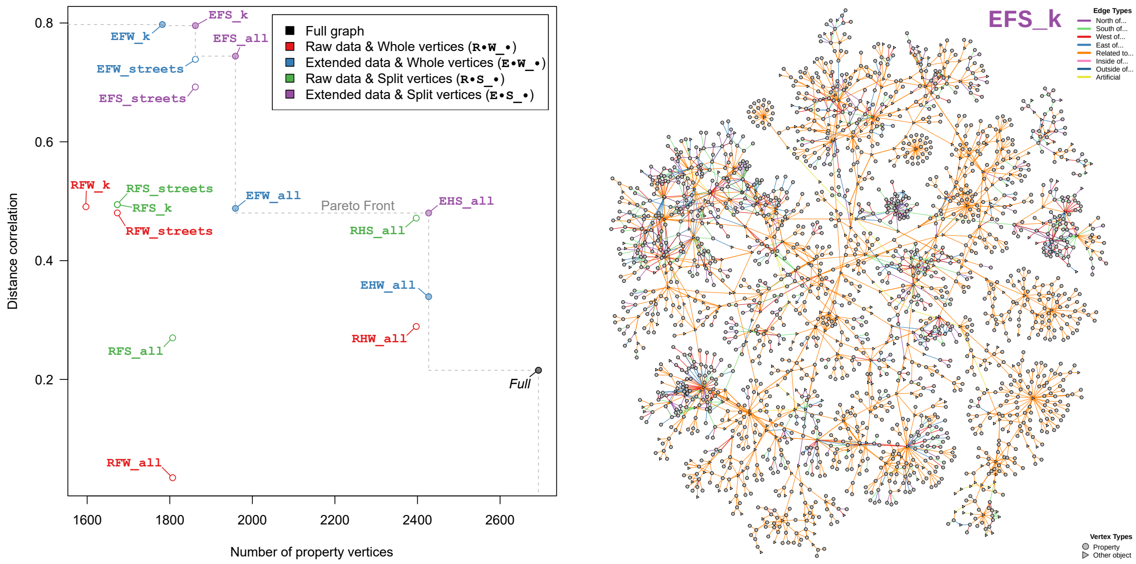

Figure 7: **Left:** Comparison of the methods from Table 3 based on coverage (number of vertices representing properties) and reliability (distance correlation). **Right:** best confront graph according to our criteria, shown using the Yifan-Hu layout [22]. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

of this graph at a meso scale, i.e. from the perspective of its community structure. Our goal here is to partition the confront network into spatial zones representative of the tenants' perception and of the use of urban space. For this purpose, we rely on the traditional modularity function [28] to assess the quality of the partition, but we also consider more subjective aspects, in order to take into account our objective of interpreting the detected communities. First, we want to get a reasonable number of communities, as it would not be practically feasible to interpret hundreds of spatial segments. Second, we want the community size to be relatively uniform: getting a single giant community and many very small ones would be of any help.

We assess seven standard community detection algorithms implemented in the `igraph` [11] R library: Edge-betweenness [28], FastGreedy [9], Leading Eigenvectors [27], LabelPropagation [31], WalkTrap [30], InfoMap [33], and Louvain [6]. According to our criteria, the Louvain method provides the most relevant partition of the graph. It reaches a modularity of 0.93, and contains 31 communities, whose main characteristics are described in Table 4. Each statistic is computed by considering only the subgraph induced by a community. Community size ranges from 13 to 168 vertices, including 67% to 91% of properties. The average distance is relatively homogeneous, ranging from 1.93 to 4.35, but the diameter is more heterogeneous, ranging from 3 to 13. The distance correlation is very heterogeneous, ranging from 0.03 to 0.78, showing that space is not modeled in all communities with the same reliability.

In the following, in order to illustrate the historical relevance of the communities detected in our confront network, we discuss a few particularly interesting examples, both in terms of how they are distributed over the city space (Section 7.1), and of the types of spatial objects that constitute them (Section 7.2).

7.1 Organization of the Communities

The spatial distribution of communities generally aligns with the logic of parochial territory, as shown in Figure 8. However, while some communities consist of properties that depend on a single parish, others are quite revealing of uncertain boundaries or the interweaving of parishes. One community, which spreads through the heart of the city, comprises properties associated with five different parishes (C18), while three communities include properties from three distinct parishes (C8, C10, and C17), and nine from two parishes. See Figure 17 in Appendix E for further details.

Here, it is not the parish that defines the neighborhood, but primarily the spatial proximity and

Table 4: Main topological characteristics of the communities identified in the selected spatial graph: numbers of vertices (n) and edges (m), density (δ), number of property vertices and proportion of such vertices relative to the community (*Properties*), diameter ($d_{\max}$), harmonic mean of the graph distance ($\langle d \rangle$), Spearman's correlation between the graph and spatial distances (ρ_d).

Com.	n	m	δ	# Properties		$d_{\max}$	$\langle d \rangle$	ρ_d
1	156	429	0.0177	132	84.62%	11	3.78	0.75
2	89	239	0.0305	73	82.02%	8	2.92	0.52
3	90	160	0.0200	62	68.89%	10	3.60	0.50
4	20	24	0.0632	16	80.00%	6	2.27	0.32
5	163	449	0.0170	132	80.98%	9	3.55	0.57
6	31	70	0.0753	24	77.42%	6	2.28	0.56
7	52	82	0.0309	42	80.77%	10	3.22	0.36
8	126	283	0.0180	105	83.33%	8	2.97	0.34
9	168	278	0.0099	135	80.36%	10	3.79	0.48
10	41	54	0.0329	32	78.05%	7	3.04	0.50
11	33	49	0.0464	29	87.88%	6	2.63	0.21
12	132	210	0.0121	105	79.55%	9	3.72	0.53
13	31	34	0.0366	27	87.10%	3	1.93	0.27
14	61	100	0.0273	55	90.16%	7	2.91	0.22
15	115	144	0.0110	105	91.30%	9	3.56	0.03
16	55	56	0.0189	39	70.91%	13	3.89	0.78
17	120	154	0.0108	90	75.00%	12	4.35	0.61
18	91	120	0.0147	78	85.71%	10	3.88	0.75
19	99	148	0.0153	89	89.90%	10	3.57	0.38
20	28	35	0.0463	23	82.14%	8	2.69	0.11
21	91	111	0.0136	69	75.82%	12	4.21	0.52
22	106	153	0.0137	95	89.62%	6	3.09	0.29
23	96	179	0.0196	86	89.58%	7	3.26	0.48
24	13	21	0.1346	11	84.62%	5	1.97	0.31
25	49	68	0.0289	33	67.35%	13	3.62	0.24
26	26	27	0.0415	20	76.92%	7	2.93	0.60
27	28	30	0.0397	20	71.43%	9	3.14	0.36
28	40	136	0.0872	36	90.00%	5	1.97	0.54
29	21	21	0.0500	16	76.19%	8	2.77	0.63
30	68	126	0.0277	46	67.65%	10	3.24	0.63
31	55	91	0.0306	37	67.27%	10	3.41	0.32

what it implies. Parish boundaries do not segment the perceived space of the tenants. They are not physically visible in the city, which partly explains why they do not reflect in the spatial distribution of some communities. These boundaries are inevitably less significant than a very concrete and physical division of urban space, such as the old city walls, even if they have disappeared for several generations. In this sense, the mental division of urban space, inherited from the old city delineation, is quite perceptible in the structure of the detected communities.

Out of the 31 communities, only five include properties located both within the old intramuros and beyond the old city walls (C17, C21, C25, C29, C31), as shown in Figure 9. These are located at the border of the two spaces. The perception of individuals does not depend on a pre-established order based on the official division of the city. It is primarily founded on the spatial proximity of properties and, by extension, of individuals. More than its relation to the administrative subdivision of the city, the division highlighted by the communities aligns with that drawn by the oldest and most structuring streets of the city, constrained by the very topography of the places and inherited spaces. Thus, the distribution of communities is constrained by certain elements that concretely segment the space: the topography with the Rock of the Doms, the canals, and the old walls. These elements clearly separate properties located near each other. When they are not connected by bridges, they hinder communication and, consequently, sociability between

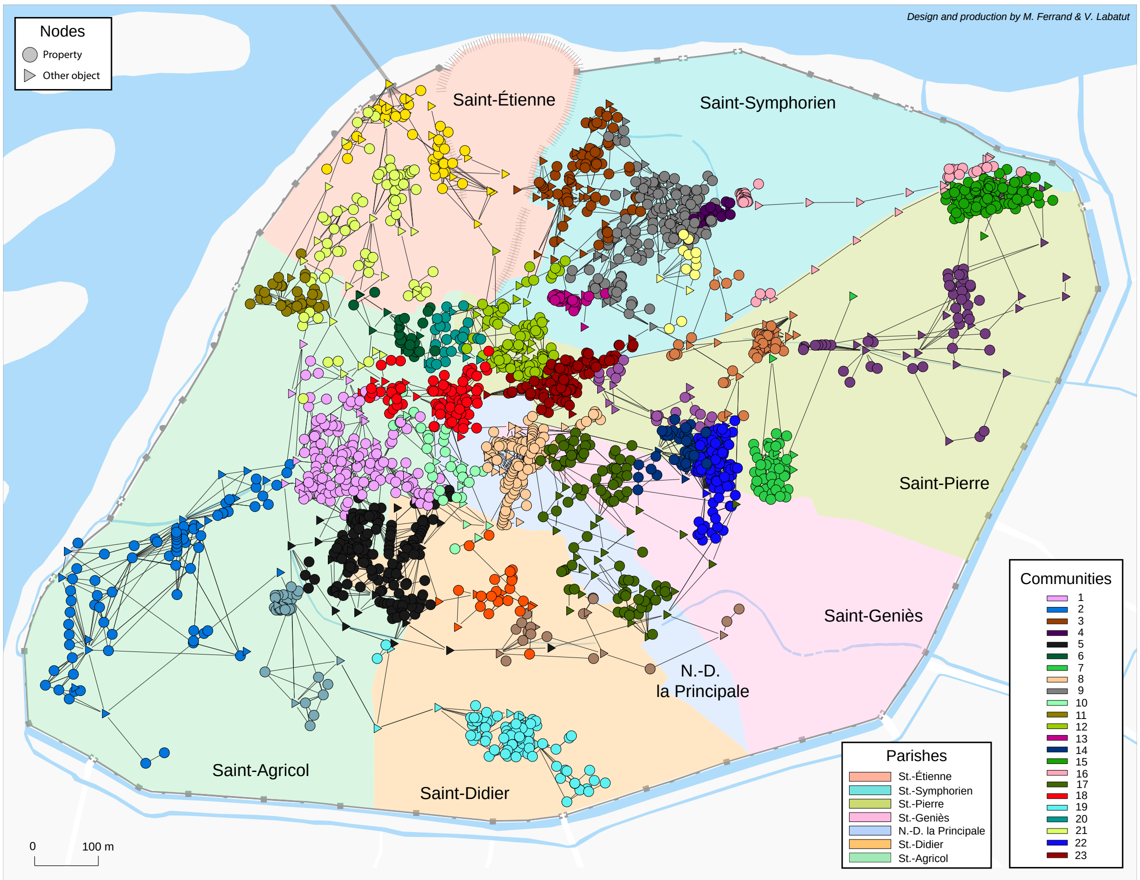

Figure 8: Distribution of the communities over the 7 historical parishes of Avignon. Each vertex color represents a specific community. The vertices are placed depending on their spatial position (known or estimated). Figure available at 10.5281/zenodo.14175830 under CC-BY license.

individuals.

Beyond spatial proximity, it is also the interactions of individuals around common places that truly create the community, in this case, the neighborhood.

7.2 Composition of the Communities

The communities exhibit a wide diversity of composition. As shown in Table 4 and illustrated in Figure 10, they contain mainly properties, whose characteristics may vary from one neighborhood to another, and invariants — more or less numerous and attractive depending on the case. Different rationales of construction can be identified, that are primarily related to the uses of space, their evolution, or, on the contrary, their persistence and stability.

7.2.1 Within the Old Walls

Depending on the neighborhoods, certain elements are predominantly used by the owners to locate and identify their properties, occupying a central place in the construction of individuals' lived spaces. In the old intra-muros of the city, that of the 13$^{\text{th}}$ century, it is often the main roads of the city, those that have structured the city since antiquity. These are main arteries that traverse the neighborhood, punctuating exchanges and social interactions. Many communities are built around these streets, though not all of these communities have the same characteristics. While some communities are more residential, others are resolutely oriented towards craftsmanship and commerce, or even towards accommodating travelers. The places that unite them are often revealing of their typology.

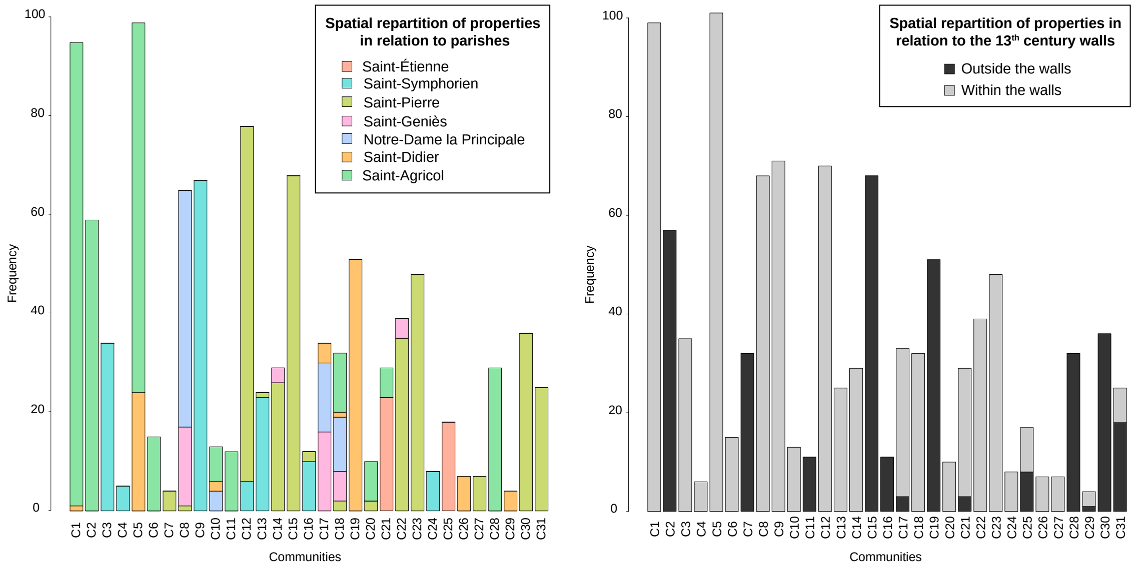

Figure 9: Distribution of property location over communities, in terms of parochial membership (left) and relative to the old walls (right). Figure available at 10.5281/zenodo.14175830 under CC-BY license.

Given its location near the Pont Saint-Bénézet and the Rhône river, C21, for example, has many inns, which are significant social places and important topographical landmarks, often with very conspicuous signs. Like the inns, the bathhouses are places of social interaction and notable landmarks in a neighborhood. Alongside the most structuring streets of the community and the inns, it is thus a bathhouse that holds one of the highest degrees in this community.

Neighborhoods with a much more residential aspect than the one we just mentioned also aggregate around many roads. Community C5 is a good example. Rue des Ortolans plays an important role in the life of the neighborhood and in the perception the owners have of it. It has long been dominated by the presence of a prominent Avignonese family, who gave its name to the street and who settled in the city well before the arrival of the popes. The residential character of this neighborhood is notably based on the fact that it is a recently constructed area, at least more recent than the heart of the city further north.

A bit further north of C5, community detection reveals a particularly imposing neighborhood both in the number of nodes it contains and its extent (C1). It has an obvious residential character and includes notably beautiful and large residences. Until 1370, one house is particularly used by contemporaries to situate themselves in this area: it is that of Tymburge Vayrane, a property owner with a rich real estate portfolio throughout the city. This house is mentioned as a landmark more often than some streets. This illustrates how much the neighborhood and social interactions directly influence the perception of the users.

Like the ones we just mentioned, C8 is a very good example of the centrality of streets in the composition and definition of neighborhoods. The community has a relatively small diameter (8) given the number of its nodes (126). Spatially speaking, it is particularly dense, which is quite revealing of the very genesis of the neighborhood and its development from the 11$^{\text{th}}$ century. The extent of this community corresponds very likely to that of the bishop's boroughs called Scofaria and Pelliparia. The regularity of the plots around the churches, especially to the south of the main street, suggests that the bishop, holder of the borough, had certainly intervened very precisely in the subdivision of certain parts of the area.

Community C18 has a very different characteristic, much less residential. It is located in the heart of the city, spanning several parishes. It is built around the most important and oldest commercial areas and marketplaces of the city. This neighborhood is resolutely commercial, and this is particularly evident in the profile of the individuals one encounters and the types of properties they own. Thus, in addition to their houses, the owners declare numerous shops or stalls, two types

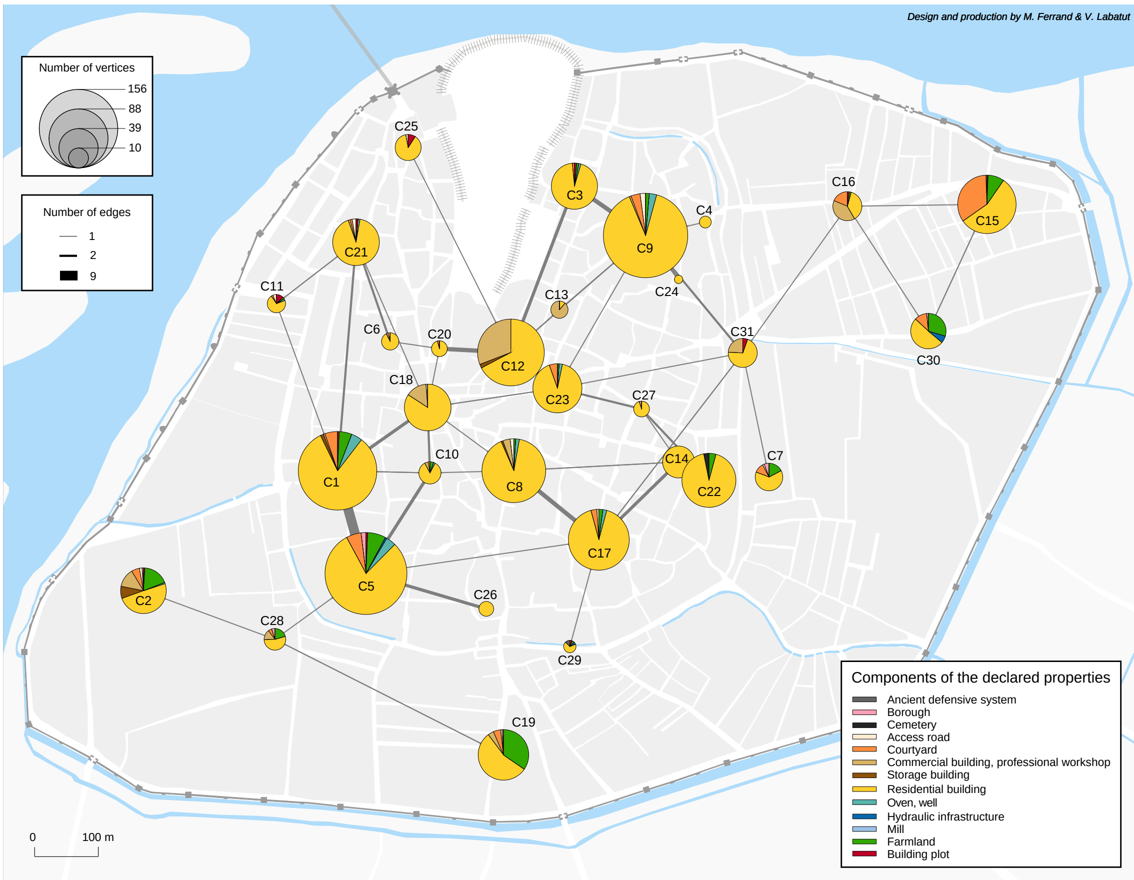

Figure 10: Simplified representation of the communities: each node represents a community from the original network. Node size and link width are proportional to the numbers of vertices constituting the original communities, and to the number of edges between them in the original graph, respectively. The pie charts show the composition of the communities in terms of types of spatial objects. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

of properties not found in more residential neighborhoods like C5. Conversely, certain features of residences found in other neighborhoods are completely absent here: there are no gardens or courtyards; the space is primarily used by and for commerce.

If the streets are at the heart of neighborhood construction, it is primarily because they are the sites of social interactions. These streets are organized around a significant and unifying household, host commercial and artisanal activities, and are central to the city's political representation. Each neighborhood also has essential elements for the daily lives of individuals, such as wells, ovens, and cemeteries, which contribute to the construction of individuals' perception of places. These are undeniable landmarks of the urban landscape, places of sociability, exchanges, and meetings. As such, they are often the vertices with the highest degree in communities. In the city, burial sites are thus privileged and central spaces for exchanges and sociability. They must also be associated with public places such as streets or even more so, squares. Community 13 is a very good example. It is very small and particularly concentrated (31 vertices for a diameter of 3). All these properties are connected to the Saint-Symphorien cemetery. The degree of centralization of the community is, in this respect, quite high. Here, the cemetery is the spatial reference par excellence, and the neighborhood is entirely built around the burial site.

7.2.2 Beyond the Old Walls

Within the old walls, communities are thus constructed by the proximity relationship between properties and individuals and by the presence of topographical landmarks that shape the perception and usage of individuals in the neighborhood. The neighborhoods located outside the walls of the 13^{th} century follow very diverse logics, which differ in several ways from the communities studied so far.

The largest neighborhood detected in this area corresponds to C2. At first glance, it might seem to follow the same construction logics as those located within the old walls. However, it has certain characteristics specific to its location and its distance from the historical center of the city. To locate their properties, the owners refer not only to the streets but also to recent urbanization: the numerous boroughs developed in this area on former agricultural land. Alongside these boroughs, many gardens are still present. People frequently use these gardens as reference points. This neighborhood has another unique characteristic: in addition to its semi-rural aspect, created by the number of gardens adjoining the new houses of the boroughs, it houses the dependencies of many services from the old intra-muros. It is likely that this community also hosts the pleasure homes of certain prelates, as suggested by the number of residences described as large and beautiful, accompanied by gardens, courtyards, or wells. This neighborhood also has a significant presence of artisans and merchants. The presence of large gardens still cultivated at the end of the 14^{th} century must have given a rural appearance to the place, sharply contrasting with the old intra-muros, which was very densely built and populated. As such, it is overall not densely built. Two other communities (C19 and C30), located outside the 13extsuperscriptth century walls, share characteristics with C2 and could thus be described as semi-rural neighborhoods. They retain the marks of significant rurality while being partly occupied by recent urbanization in certain areas.

Finally, a specific typology of community outside the 13extsuperscriptth century walls must be highlighted. These are what could be called the borough communities, such as C15. Here, the borough forms the neighborhood. This is quite noticeable when looking at the elements to which individuals refer to situate themselves in this space. Firstly, it is the straight street of the borough that is used: it unites the community built around it, suggesting the planning of the road in the borough's subdivision project. Although less used, another street is mentioned; it is called the "vicinal street" of the borough; there is also a cross street, a dead-end, and a square. In all cases, the attachment of these elements to the borough is specified. These details remind us that while these streets are not structuring at the city level, they are strongly so at the neighborhood and borough level. Whereas the boundaries of the previously studied neighborhoods were more difficult to perceive, especially for those within the 13^{th} century walls, here the neighborhood space is clearly delineated. It is no longer just a perceived, subjective space with fuzzy boundaries, constructed by the proximity and sociability of individuals around various places and inherited spaces. In the case of the "borough neighborhood," the boundaries are tangible and inherent to the very genesis of the place. Proximity and sociability are indeed present, as evidenced by the presence of a large borough well to which the owners sometimes refer. However, the neighborhood is built here around a predefined space, not an unplanned process inherited from an "already existing space." The typology of the declared properties once again highlights this, forming a particularly homogeneous neighborhood.

The graph extracted using our method is a sufficiently good approximation of spatial distance for the communities to correspond to spatial segments representative of real neighborhoods in the sense of lived space. In this regard, the detection of communities in the graph shows a partition of the city consistent with social practices. The elements that define the communities are multiple and complementary (livability, buildings, neighborhoods, commerce) and they are essentially places of sociability or symbolic markers of the urban landscape. They pace and define spatial perceptions. These perceptions are primarily based on the layering of heritages, with evolutions being integrated only slowly.

8 Conclusion

In this article, we tackled the problem of segmenting a historical urban space based on incomplete data. We focused on the specific case of medieval Avignon, during the papacy. On the one hand, historical sources rarely provide the exact locations of the buildings constituting the city, but on the other hand, they describe how they are located relative to each other. We adopted a graph-based approach to take advantage of this relational information while modeling the city, and proposed several methods to extract such confront networks. We designed two conflicting objective criteria to select the optimal extraction method: data coverage and distance correlation. It turns out the best method requires discarding some of the information provided by the original historical sources in order to preserve the spatial proximity encoded in the graph. Moreover, our results show that leveraging certain information from secondary sources can considerably improve the quality of the graph. Finally, we partitioned the best confront network through community detection, in order to perform a qualitative analysis of the resulting segmentation of the urban space. Our discussion showed that the communities are completely consistent with historical knowledge, and help understand how people perceived urban space at the time.

We identify mainly two perspectives to extend our work. First, the land registries that we used in this article are very common in historical studies, especially during the Medieval Ages. However, they do not always take exactly the same form, and can also vary in content from one city to the other, depending on the local practices. We want to assess the robustness of our extraction approach by applying it to other cities, which will require some adaptation. For example, in the case of the city of Orleans [14], the spatial relationships partly differ from those described in Section 3.3 and Table 6 for Avignon: there are only very few cardinal relationships. But this could be compensated by leveraging other types of information, for instance regarding the notaries and families involved, and their social relationships. The second perspective concerns the estimation of the missing absolute positions. In the case of Avignon, the position of certain spatial objects is known with reasonable certainty, and can be leveraged together with the spatial relationships encoded in the confront networks, to interpolate the position of the other objects. A simple baseline consists in averaging the position of the neighboring objects, but this does not account for the semantics of the spatial relationships. We want to take advantage of Graph Neural Networks [17] to perform a better prediction, possibly also using the many vertex attributes at our disposal.

Resource Availability

The source code written to implement our extraction methods and experiments is available online at https://github.com/CompNet/MedievalAvignon. The input data are included in this repository, whereas all the files produced during their processing are available at https://doi.org/10.5281/zenodo.14175830.

Acknowledgment

This research benefited from the support of the *Agorantic* interdisciplinary research federation (FR 3621), through the funding of Margot Ferrand's PhD. thesis and of the *HistoGraph* research project.

References

[1] J.-L. Arnaud. *Analyse spatiale, cartographie et histoire urbaine.* Parcours méditerranéen. Parenthèses/MMSH, Marseille/Aix-en-Provence, FR, 2008. URL https://www.editionsparentheses.com/analyse-spatiale-cartographie-et.

[2] M. Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011. doi:10.1016/j.physrep.2010.11.002.

[3] P. Bertrand. *Les écritures ordinaires : sociologie d'un temps de révolution documentaire (entre royaume de France et Empire, 1250–1350).* Histoire ancienne et médiévale. Publications de la Sorbonne, Paris, FR, 2015. URL http://www.editionsdelasorbonne.fr/fr/livre/?GCOI=28405100607730.

[4] N. Blagus, L. Šubelj, and M. Bajec. Self-similar scaling of density in complex real-worldnetworks. *Physica A*, 391(8):2794–2802, 2012. doi:10.1016/j.physa.2011.12.055.

[5] M. Bloch. Les plans parcellaires en France. *Annales d'histoire économique et sociale*, 1(1): 60–70, 1929. doi:10.3406/ahess.1929.1039.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008(10):P10008, 2008. doi:10.1088/1742-5468/2008/10/P10008.

[7] S. P. Borgatti and M. G. Everett. Network analysis of 2-mode data. *Social Networks*, 19(3): 243–269, 1997. doi:10.1016/S0378-8733(96)00301-2.

[8] G. Brunel, O. Guyotjeannin, and J.-M. Moriceau, editors. *Terriers et plans-terriers du XIIe au XVIIIe siècle : Actes du colloque de Paris*, Paris, FR, 1998. Association d'Histoire des Sociétés Rurales / École Nationale des Chartes. URL `https://books.google.fr/?id=mjRpmXRfNkYC`.

[9] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004. doi:10.1103/PhysRevE.70.066111.

[10] A. Claveirole and É. Pélaquier, editors. *Le compoix et ses usages – Actes du colloque de Nîmes*, Montpellier, FR, 1999. Publications de l'Université Paul Valéry-Montpellier 3. URL `https://books.google.fr/?id=JNO8bwAACAAJ`.

[11] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, 1695(Complex Systems), 2006. URL `http://igraph.sf.net`.

[12] E. Estrada. *The Structure of complex networks: Theory and applications*. Oxford University Press, 2011. doi:10.1093/acprof:oso/9780199591756.001.0001.

[13] M. Ferrand. *Usages et représentations de l'espace urbain médiéval : Approche interdisciplinaire et exploration de données géo-historiques d'Avignon à la fin du Moyen Âge*. Phd thesis, Avignon Université, 2022. URL `https://www.theses.fr/2022AVIG1002`.

[14] K. Fianu. L'utilisation des actes d'après les registres notariés orléanais du xve siècle. *Cahiers de recherches médiévales et humanistes*, 22:457–479, 2011. doi:10.4000/crm.12625.

[15] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. doi:10.1016/j.physrep.2009.11.002.

[16] R. Fossier. *Polyptiques et censiers*. Brepols, Turnhout, BE, 1978. URL `https://bibliothequearchives.seinemaritime.fr/Default/doc/SYRACUSE/243138/polyptyques-et-censiers-par-robert-fossier`.

[17] William L. Hamilton. *Graph Representation Learning*, volume 46 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool, 2020. doi:10.2200/s01045ed1v01y202009aim046.

[18] F. Hautefeuille. Géolocalisation des sources fiscales pré-révolutionnaires : la quadrature du cercle. *Bulletin du Centre d'Études Médiévales d'Auxerre*, HS-9, 2016. doi:10.4000/cem.13800.

[19] A.-M. Hayez. Les bourgs avignonnais au XIVe siècle. *Bulletin philologique et historique du Comité des Travaux Historiques et Scientifiques*, pages 77–102, 1977. URL `https://bibliotheques.avignon.fr/in/faces/details.xhtml?id=p::usmarcdef_0000359794`.

[20] A.-M. Hayez. *Le terrier avignonnais de l'évêque Anglic Grimoard (1366-1368) rédigé par Sicard de Fraisse*. Comité des Travaux Historiques et Scientifiques, Paris, FR, 1993. URL `https://gallica.bnf.fr/ark:/12148/bpt6k6431727c.texteImage`.

[21] A.-M. Hayez. Livrées avignonnaises de la période pontificale. *Mémoires de l'Académie de Vaucluse*, 8(3):33–89, 1993. URL `https://bibliotheques.avignon.fr/in/faces/details.xhtml?id=p::usmarcdef_0000119299`.

[22] Y. Hu. Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, 10:37–71, 2006. URL http://yifanhu.net/PUB/graph_draw.pdf.

[23] M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014. doi:10.1093/comnet/cnu016.

[24] J. Levy and M. Lussault. *Dictionnaire de la géographie et de l'espace des sociétés*. Belin, Paris, FR, 2003. URL https://shs.hal.science/halshs-01252959.

[25] F. Mazel. Pouvoir comtal et territoire : réflexion sur les partages de l'ancien comté de provence au XIIe siècle. *Mélanges de l'École Française de Rome*, 123(2):467–48, 2011. doi:10.4000/mefrm.634.

[26] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45: 167–256, 2003. doi:10.1137/S003614450342480.

[27] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006. doi:10.1103/PhysRevE.74.036104.

[28] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. doi:10.1103/PhysRevE.69.026113.

[29] P. Pansier. *Dictionnaire des anciennes rues d'Avignon*. Roumanille, Avignon, FR, 1930. URL https://bibliotheques.avignon.fr/in/faces/details.xhtml?id=p::usmarcdef_0000031742.

[30] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006. doi:10.7155/jgaa.00124.

[31] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007. doi:10.1103/PhysRevE.76.036106.

[32] X. Rodier, M. Le Couédic, B. Jouve, F. Hautefeuille, S. Leturcq, and E. Fieux. De l'espace aux graphes. mesurer les dynamiques spatiales des terroirs villageois. In *XLIIIe Congrès de la Société des historiens médiévistes de l'Enseignement supérieur public*, pages 99–118. Éditions de la Sorbonne, 2012. URL https://books.openedition.org/psorbonne/28550.

[33] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118, 2008. doi:10.1073/pnas.0706851105.

[34] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge, US, 1994. doi:10.1017/CBO9780511815478.

A Historical Sources

Table 5 lists of all the documents used as primary historical sources in the present work. The conservation reference numbers include the following acronyms:

- AAV: Apostolic Archives of the Vatican;
- AMA: Avignon Municipal Archives;
- DABdR: Departmental Archives of Bouches-du-Rhône;
- DAV: Departmental Archives of Vaucluse;
- PACB: Private Archives of the Château de Barbentane.

The Accounts of the City's Clavigers lists the properties belonging to the Pope since 1348.

Table 5: List of documents used as the main historical sources in the present work. See the main text for the meaning of the acronyms used in column *CRN* (Conservation Reference Numbers).

Document	**Date**	**CRN**	**Language**
Terrier of Bishop Anglic Grimoard	1366	DAV, 1G10	Latin
Terrier of the Provost of the Cathedral	1366	DAV, G536	Latin
Terrier of Sainte-Catherine	1366	DAV, 71H5	Provençal
Terrier of the Urban Community	1362	AMA	Latin
Terrier of the Repenties of Notre-Dame des Miracles	1395	DAV, 107H15	Latin
Terrier of the Collegiate Chapter of Saint-Pierre	1380	DAV, 9G2	Latin
Accounts of the City's Clavigers	1384	AAV	Latin
Terrier of the Hospital of Saint-John of Jerusalem	1350	DABdR	Provençal
Terrier of the Cabassoles	1319	PACB	Latin

B Relationship Normalization

Table 6 lists the relationships between objects that appear in the historical sources, with their English translation, and how we normalize them to ease graph extraction.

For certain relationships, the normalization depends on whether the considered object is two-dimensional, by opposition to punctual (building, shop...) and linear (streets, channels...) objects. Two-dimensional objects include boroughs, parishes, and cardinal's liveries.

Relationship *Egal* (i.e. *Same as*) is particular, as it does not appear explicitly in the land registers, but is inferred manually. It connects two surface forms representing the same real-world entity: we merge them during the normalization.

C Extracted Graphs

Figure 11 shows the full graph, i.e. the graph extracted without discarding any vertex or edge, and without inserting any additional relationship. Vertex location is determined using the Yifan-Hu layouting algorithm [22].

Figures 12 and 13 show the graphs extracted through all the optional steps considered in the article. The layout corresponds to the spatial position of the vertices: exact when it is known, and estimated when it is unknown.

D Extraction Statistics

D.1 Long Streets

Figure 14 shows the plots used to select the optimal value of parameter k during the extraction process of graphs `RFW_k` (*raw* data, flat relations, whole vertices, top k streets removed) and `EFW_k` (same thing, but with *extended* data). This parameter corresponds to the number of streets removed during the extraction process. The removal starts with the longest streets (cf. Section 5.2.1). The y axis corresponds to Spearman's correlation between the graph and spatial distances. The

Table 6: Types of relationships appearing in the land registers to position properties relatively to other spatial objects, their translation in English, and how they are normalized before graph extraction.

Occurrences	**Relationship**	**Translation**	**Normalized Form**
73	*Iuxta*	Near...	Related to...
250	*Juxta*		
82	*Prope*		
2	*Proxime*		
3	*In Angulo*	At the corner of...	Street: Related to...
1	*In Cantono*		2D objects: Inside of...
13	*In Compito Sive Cantono*		
8	*In Introytu*	At the entrance of...	Street: Related to...
			Others: Inside of...
91	*Extra*	Outside of...	Outside of...
2,009	*In*	Inside of...	2D objects: Inside of...
4	*Intra*		Others: Related to...
2	*Ab Opposito*	Opposite to...	Related to...
1	*Ex Opposit*		
10	*In Capite*	At the beginning of...	2D objects: Inside of...
			Others: Related to...
18	*Super*	Above...	Related to...
9	*Supra*		
521	*A Orient*	West of...	West of...
535	*A Occident*	East of...	East of...
511	*A Circio*	South of...	South of...
1	*Ab Aura Recta*		
548	*A Meridie*	North of...	North of...
2	*A Una Part*	One side facing...	Related to...
1,208	Ab Una Part		
4	*A Duabus Part*	Two sides facing...	Related to...
1	*A Tribus Part*	Three sides facing...	Related to...
21	*A Parte Retro*	Rear side facing...	Related to...
17	*A Part Ante*	Front side facing...	Related to...
1	*A Part Inferiori*	Lower side facing...	Related to...
2	*A Parte Lateris*	Lateral side facing...	Related to...
5	*A Part Posteriori*	Rear side facing...	Related to...
1	*Sive Ab Una Part*	One side may face...	Related to...
1	*Conjuncto*	Adjacent to...	Related to...
15	*Contigu*		
2	*Contiguo*		
35	*Retro*	Behind...	Related to...
59	*Ante*	Before...	Related to...
3	*Egal*	Same as...	(Merge vertices)
6	*Infra*	Below...	Related to...
29	*Subtus*		
8	*Ad*	Towards...	Related to...
4	*Apud*		
13	*Versus*		

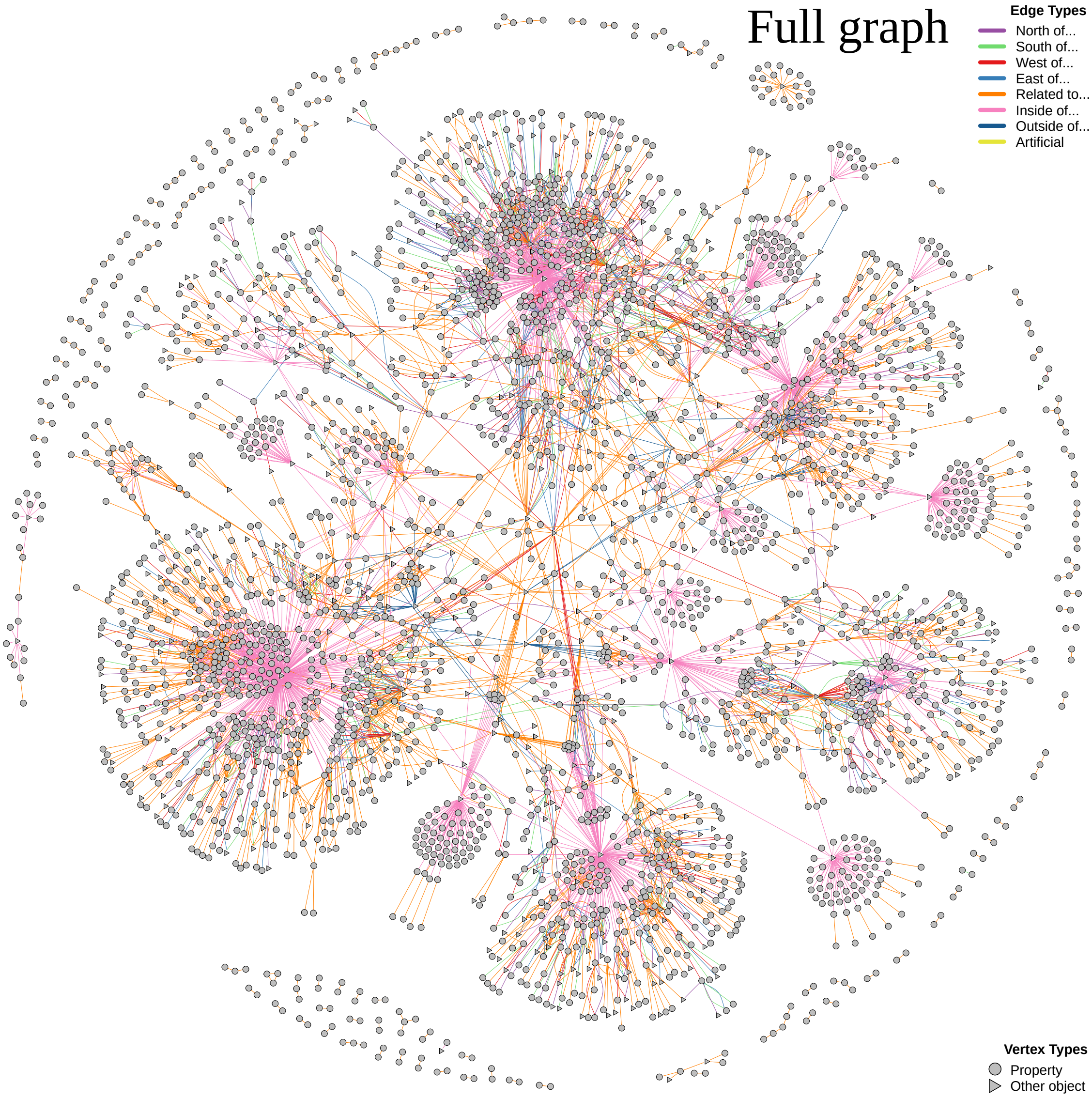

Figure 11: Full graph extracted from our database. It contains all the available raw data (but no additional relationships), without any vertex filtering. The layout is obtained using the Yifan-Hu method [22]. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

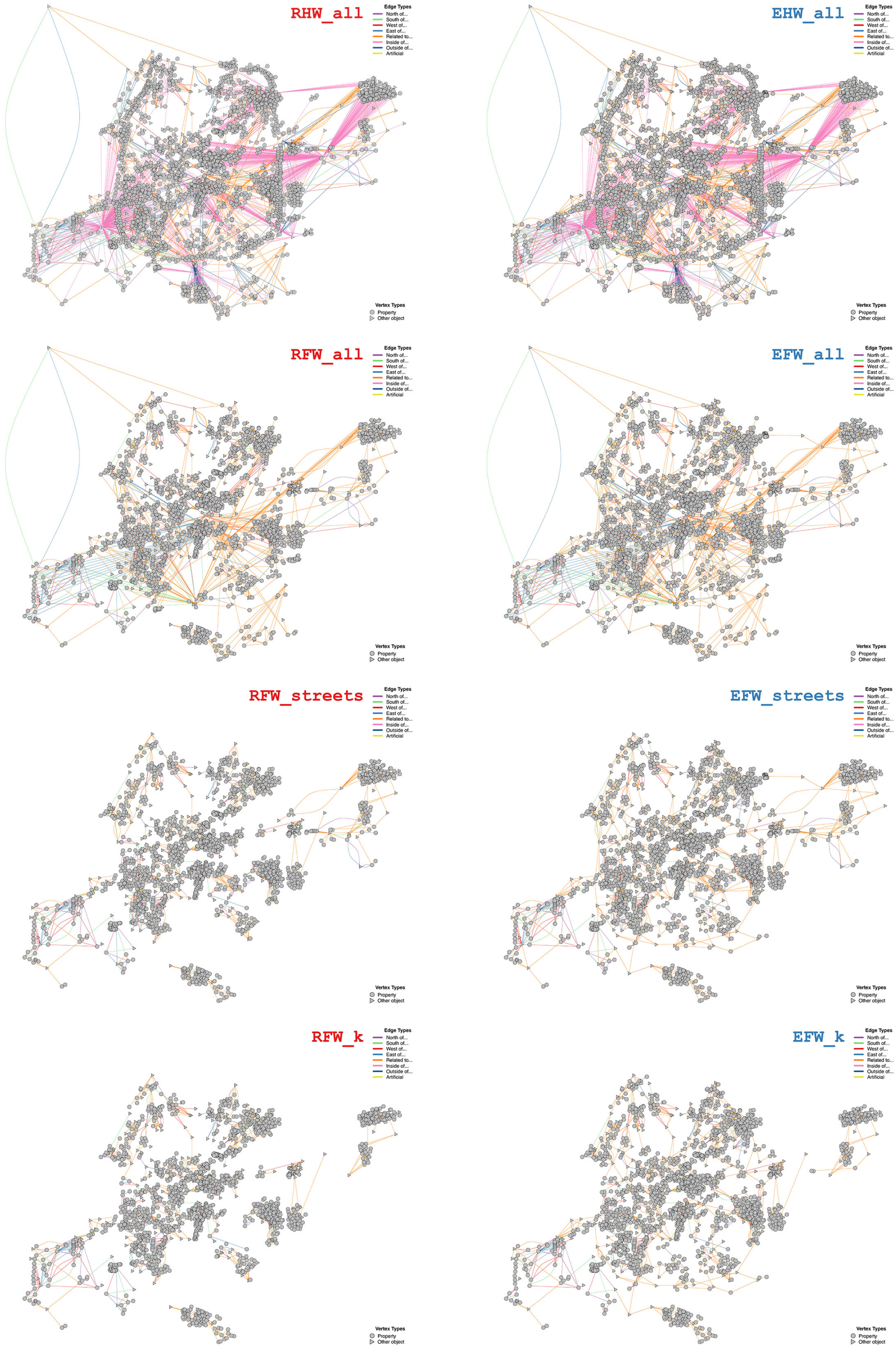

Figure 12: Graphs extracted using the whole vertex approaches (`W`). Vertices are positioned depending on their geographic location. Those whose spatial position is unknown are not shown. The left-hand column shows the graphs based on the raw data (`R`), whereas the right-hand one contains the extended versions that leverage additional relationships (`E`). The two top rows show the graphs that focus on **`all`** types of vertices, with both flat and hierarchical relations (`H`) for the first and only flat edges for the second (`F`). The two bottom rows also show flat graphs, but without the non-punctual objects except the streets (`street`) and the shorter streets (`k`), respectively. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

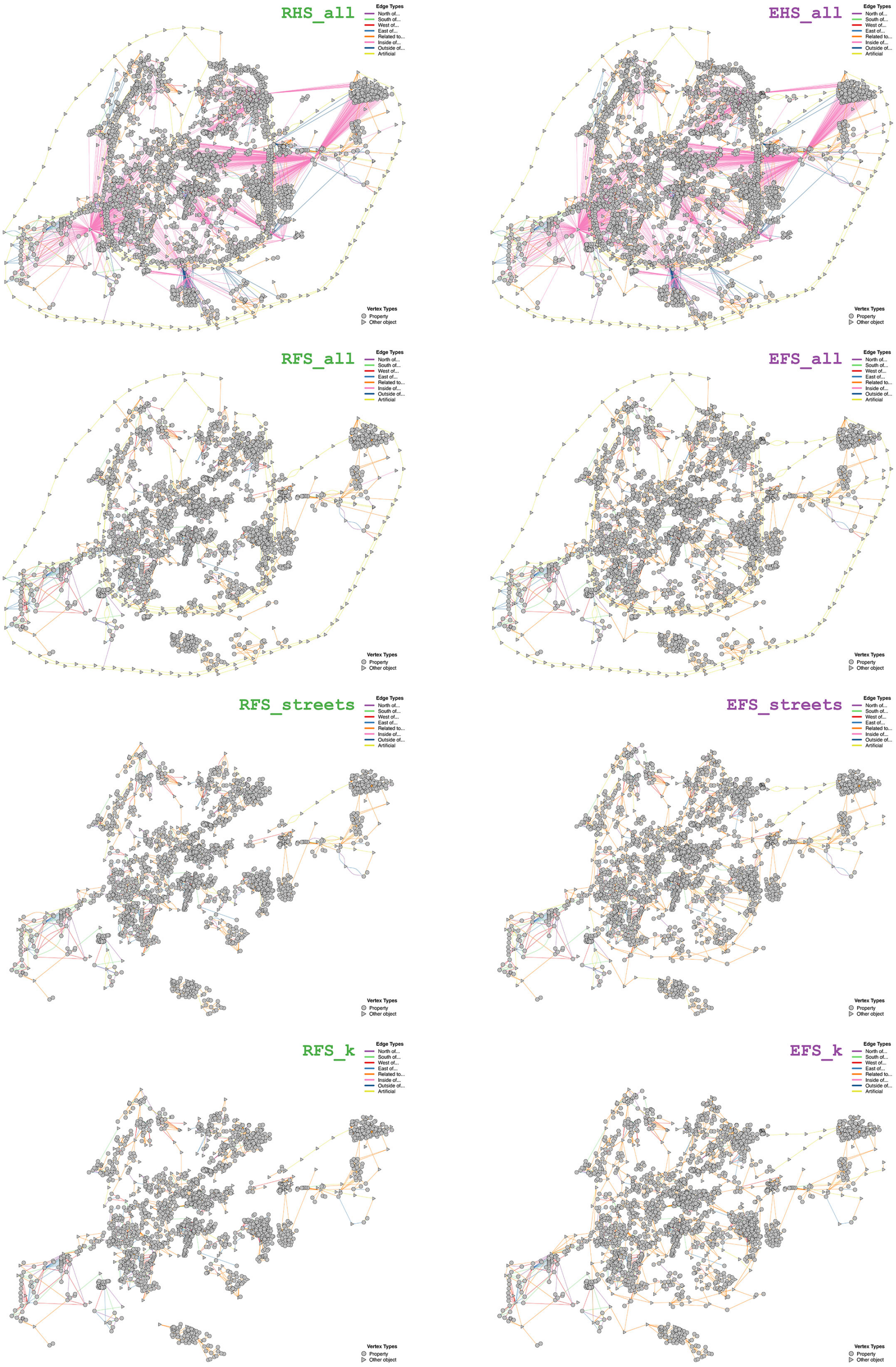

Figure 13: Graphs extracted using the split vertex approaches (`S`). Vertices are positioned depending on their geographic location. Those whose spatial position is unknown are not shown. The left-hand column shows the graphs based on the raw data (`R`), whereas the right-hand one contains the extended versions that leverage additional relationships (`E`). The two top rows show the graphs that focus on `all` types of vertices, with both flat and hierarchical relations (`H`) for the first and only flat edges for the second (`F`). All non-punctual objects are split. The two bottom rows also show flat graphs, but without the non-punctual objects except the streets (`street`) and the shorter streets (`k`), respectively, which are split. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

x axis shows the number of vertices that model properties and are still present in the graph after removing its longest streets. Each point shows both metrics for a specific value of k, indicated by its color. The dotted line materializes the Pareto front, which includes the solid dots.

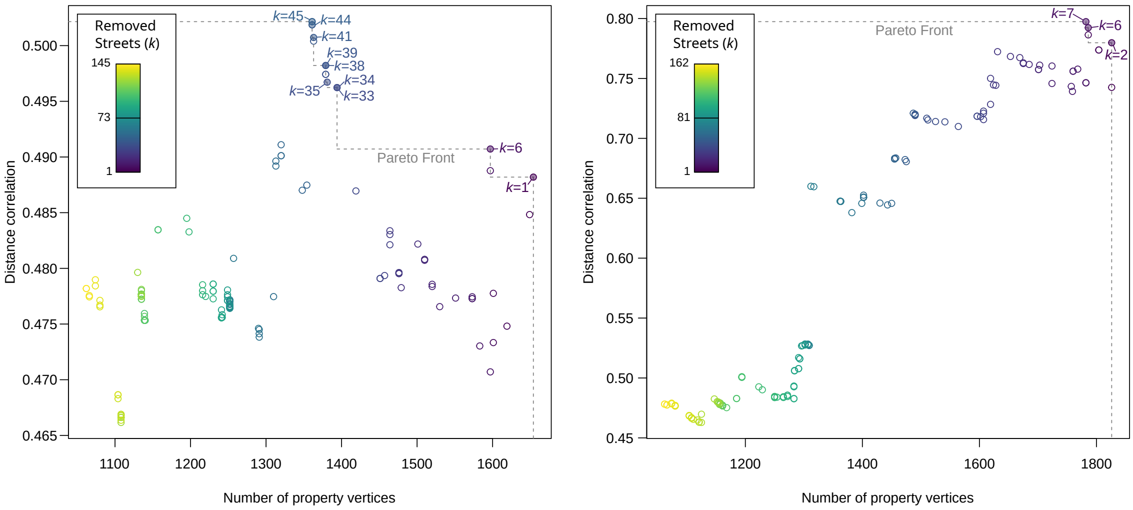

Figure 14: Selection of the optimal value of parameter k for the `RFW_k` (left) and `EFW_k` (right) extraction methods (see Section 6.3). Figure available at 10.5281/zenodo.14175830 under CC-BY license.

Figure 15 shows the evolution of our two criteria, as a function of the proportion of longest streets that undergo the removal from methods `·FW_k` (solid lines) or the splitting from methods `·FS_k` (dotted lines): on the left, the proportion of properties left in the produced graph, and on the right, Spearman's correlation between spatial and graph distances. Red lines show the results obtained on the raw data, whereas blue lines represent the extended data. Increasing k when splitting has little effect on coverage, hence the horizontal dotted lines in the left plot. Street removal (solid lines) causes the deletion of all the properties that happen to be connected only to the deleted streets, hence the decrease in coverage. For distance correlation (right plot), increasing k has not much effect when processing the raw data (red lines), be it through street removal or splitting. On the contrary, there is a clear effect when using the extended data: removing too many streets tend to decrease the distance correlation.

D.2 Distance Comparison

Figure 16 shows the comparison of the graph and spatial distances, for each graph extracted using one of the methods considered in the article. The colors are the same as in Figure 7: red for `R•W_•`, blue for `E•W_•`, green for `R•S_•`, and purple for `E•S_•`. The lines represent the means and the areas are the standard deviations.

The separated dot that appears on the right side of certain plots (ex. `RFW_all`), reveals the existence of separated components in the corresponding graph, as the graph distance between vertices located in different components is infinite. Ideally, there should be a single component, and if there are several ones, then the spatial distances matching an infinite graph distance should be the largest of all.

Visually, a good result corresponds to a relatively straight line (ex. `EFW_k`), or at least a monotonous function (ex. `RFW_k`), with low dispersion. Some plots exhibit an increasing then decreasing trend (ex. `RFS_all`), which means that large graph distances are associated with small spatial distances. This means that the graph is too linear: some shortcuts are missing to correctly encode the spatial distance.

Spanning a wide range of graph distances is also a good thing (ex. `EFS_k`), as it suggests a better "distance resolution" on the graph (i.e. a better chance that two distinct spatial distances are not associated to the same graph distance). Graphs that include too many shortcut vertices tend to exhibit a very narrow range (ex. `RHW_all`).

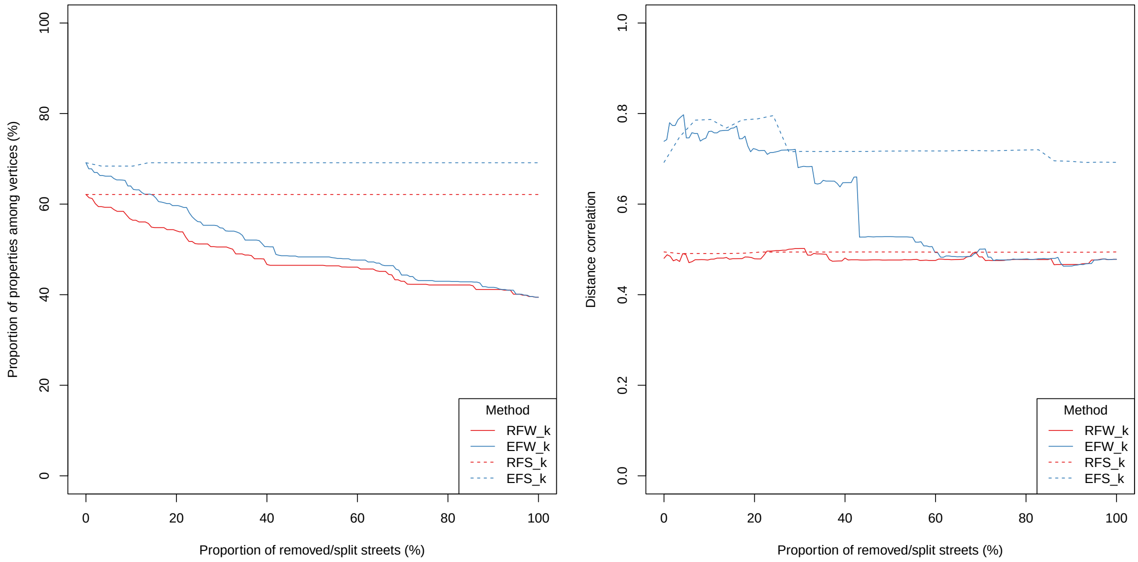

Figure 15: Evolution of the coverage (left) and distance correlation (right) as functions of the proportion of longest streets considered for removal or splitting in extraction methods ·F·_k (see Section 6.4). Figure available at 10.5281/zenodo.14175830 under CC-BY license.

E Community Structure

Figure 17 provides additional visualization of the community structures, in addition to those presented in Section 7 of the main text. It shows two so-called community networks, representing the community structure in a simplified way, where each node represents a whole community in the original graph. The left plot uses the same vertex colors as in Figure 8 to show the relative position and importance of the communities. The right plot uses the colors of the parishes, as in Figures 8 and 9, to represent the distribution of properties over parishes, for each community.

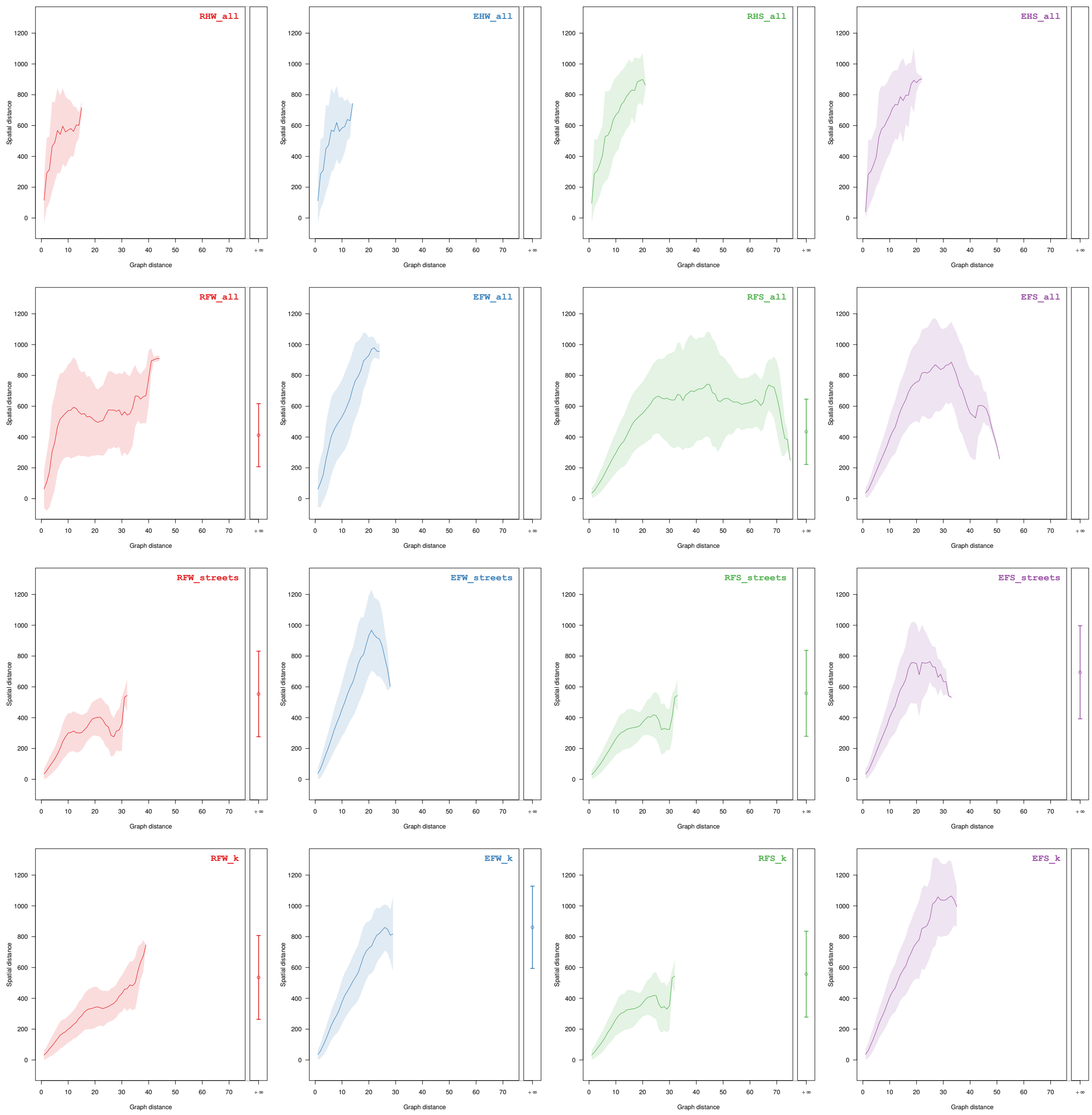

Figure 16: Comparison of graph and spatial distances, for all graphs. Figure available at 10.5281/zenodo.14175830 under CC-BY license.

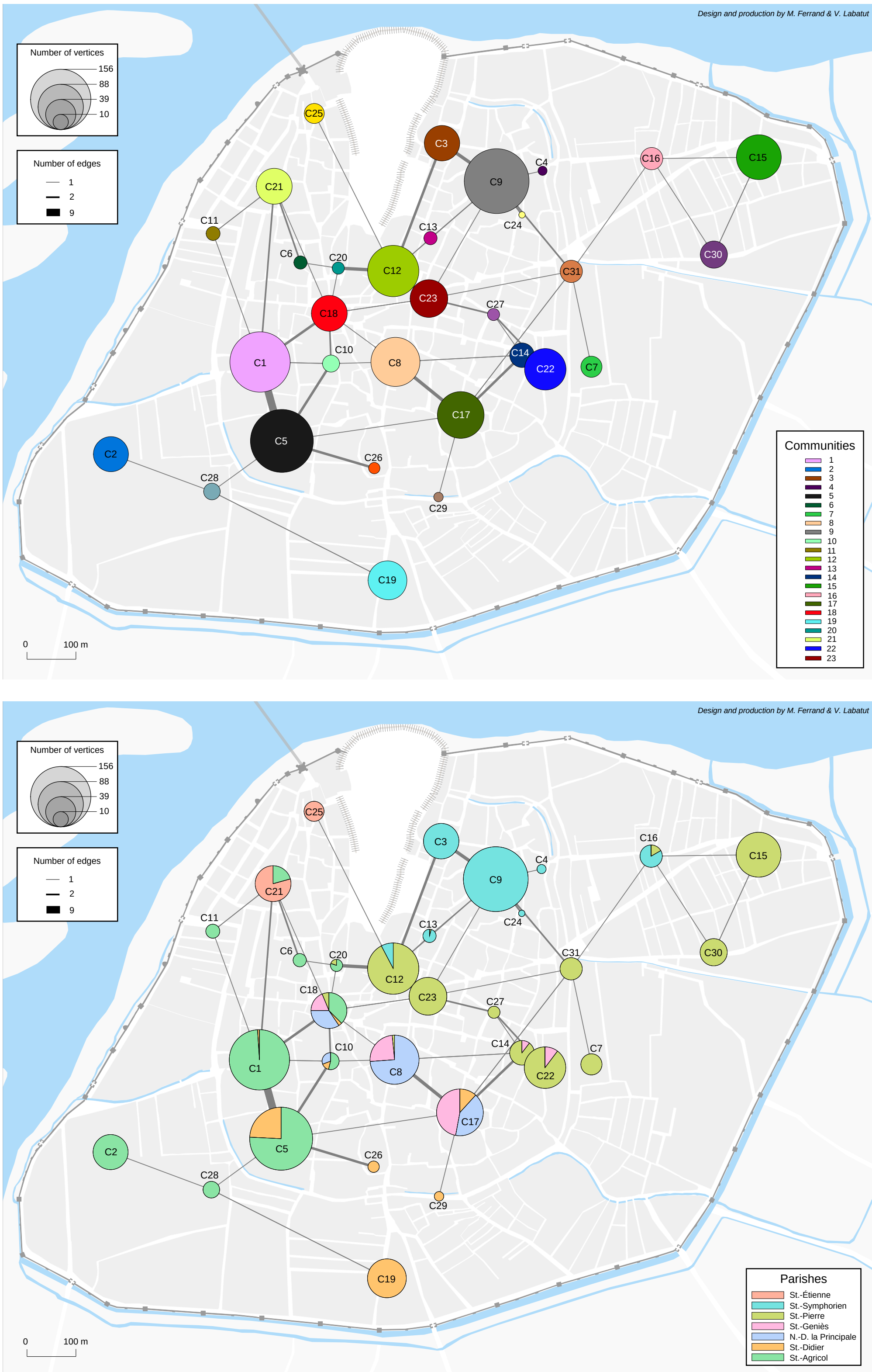

Figure 17: Two versions of the community network: each node represents a community from Section 7, its size is proportional to the number of vertices in the community, and the thickness of a link connecting two nodes corresponds to the number of edges between the matching communities in the original graph. **Top:** each community is represented using a unique color, as in Figure 8. **Bottom:** each community is represented by a pie chart showing how the parochial membership of its constituting properties is distributed. Figure available at 10.5281/zenodo.14175830 under CC-BY license.