



Article

Identification of Land Use Mix Using Point-Based Geospatial Data in Urban Areas

Mehmet Ali Akyol ^{1,*} , Tuğba Taşkaya Temizel ¹, Sebnem Duzgun ²  and Nazife Baykal ¹¹ Graduate School of Informatics, Middle East Technical University, Ankara 06800, Turkey² Mining Engineering Department, Colorado School of Mines, Golden, CO 80401, USA

* Correspondence: akyol.mehmet@metu.edu.tr

Abstract: Identifying land use mix (LUM) in urban areas is challenging, often requiring extensive human intervention and fieldwork. Accurate classification of LUM is crucial for various disciplines, including urban planning, urban economics, and public health. This study addresses this need by employing Voronoi triangulation and an entropy-based LUM formula using point-based geospatial data collected from publicly available sources. The methodology was tested in two distinct urban settings: Ankara and Kadıköy. Ankara, the capital city, provides a large and diverse urban environment, while Kadıköy, a district in Istanbul known for its dynamic urban life, offers a contrasting scenario. Results were analyzed concerning local spatial autocorrelation and point of interest (POI) intensity. The comparative analysis demonstrated that the approach performs well across different urban contexts, with improved results observed in Kadıköy due to its higher density of mixed-use development. Specifically, we managed to identify mixed land use areas with an accuracy of up to 78% and an F1-score of 83% in urban regions. These findings highlight the robustness and applicability of our approach in diverse urban environments, providing valuable insights for city planners and policymakers in optimizing the allocation of urban resources and enhancing land use efficiency.

Keywords: land use mix; point of interest data; Voronoi triangulation; local spatial autocorrelation; geospatial data mining



Citation: Akyol, M.A.; Temizel, T.T.; Duzgun, S.; Baykal N. Identification of Land Use Mix Using Point-Based Geospatial Data in Urban Areas. *Appl. Sci.* **2024**, *14*, 6871. <https://doi.org/10.3390/app14166871>

Academic Editor: Andrea L. Rizzo

Received: 8 June 2024

Revised: 30 July 2024

Accepted: 1 August 2024

Published: 6 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Land use mix, the combination of different types of land use within a given area, has become a popular topic of discussion among planners and policymakers over the years. The mix can vary in type, intensity, and spatial arrangement. Understanding land use mix (LUM) has a wide range of benefits in various areas, such as public health, transportation, and economics.

The mix of land use requires closer integration of different land functions such as residential, commercial, recreational, and many more to improve land utility for people. Knowing that a site has mixed land use is essential for a city to improve public safety, reduce commutes, and be a self-sustainable place [1]. Conventional ways of identifying LUM require tedious work with human intervention and fieldwork such as surveying and field visits [2,3].

Automating the LUM finding process using data mining techniques can reduce the effort needed. Many big data sources, some of which can aid in identifying LUM, have recently become publicly available. In particular, point of interest (POI) data have become increasingly available with the rise of location-based applications.

This study aims to identify and classify LUM in an urban area based on publicly available data sources. The study's primary motive is to model LUM in order to understand the land-use interactions of people in urban areas. We use POI data collected from ad listing websites and Google Maps API. Specifically, we aim to answer the following research questions:

- How can point-based geospatial data be used to model the LUM in urban areas?
- How effective is the combination of Voronoi triangulation and an entropy-based LUM formula in identifying and classifying mixed land use areas?
- What are the optimal POI density ranges that contribute significantly to the accuracy of LUM classification in urban areas?

Even though there are different methods to model LUM in the literature [4–7], to the best of our knowledge, identifying LUM based on POI data and Voronoi triangulation has not been investigated before in the literature. Traditional approaches have utilized travel behavior and urban form data to model LUM [4,5]. In another study, they have also contributed to this field by providing a grid-based approach by utilizing a Fishnet model to analyze land use patterns [6].

However, the innovative use of POI data combined with Voronoi triangulation to identify LUM represents a novel approach that has not been extensively explored in the literature. While traditional methods of LUM modeling have relied heavily on census data and existing land use information, the integration of POI data with Voronoi triangulation presents a promising new direction in urban planning and spatial analysis.

By integrating POI-based data with Voronoi triangulation, we aim to automate the identification of LUM, thus eliminating the need for extensive fieldwork and minimizing human intervention. This method promises to be more adaptable and capable of reflecting real-time changes in urban land use, thereby providing urban planners with a robust tool for city planning and policy-making.

1.1. Related Works

This section presents the background information on the proposed methodology and related studies. Specifically, we review land use/land cover (LULC), LULC classification, and LUM topics in general.

1.1.1. Land Use/Land Cover (LULC)

LULC refers to how people use the land and interact with the physical land type. Various authorities propose different LULC classes. CORINE, one of the widely used LULC classes, includes 44 land cover classes in a three-level hierarchy. The main categories are artificial surfaces, agricultural areas, forest and semi-natural areas, wetlands, and water bodies. The National Oceanic and Atmospheric Administration (NOAA) also categorizes land cover such as developed land, agricultural land, grassland, forest land, scrub/scrub land, bushland, palustrine wetlands, estuarine wetlands, and water and submerged land. Additionally, Sentinel, an earth observation program initiated by the European Space Agency (ESA), offers classifications, namely residential, impervious surface, agriculture, bare land, forest, and water [8].

The major limitation of these LULC classes is that they can be too specific or generic. Also, there are significant differences between the types proposed by different sources, making it subjective and problematic in particular cases. However, LULC classes provided by different entities can be useful because they provide a standardized way of categorizing land use and land cover types. Additionally, researchers, policymakers, and other stakeholders can use a common language to compare and contrast land use and land cover across different regions. For instance, the CORINE land cover classification system is widely used in Europe. The European Environmental Agency has adopted it to report on the state and evolution of Europe's environment. The NOAA's land cover classification system is used in the United States for environmental monitoring and reporting purposes.

1.1.2. LULC Classification

There has been a significant amount of research on LULC classification in the literature. These studies have focused on utilizing various data sources such as point-based data and satellite imagery to identify LULC. For instance, the study by Liu et al. [9] combined data from OpenStreetMap road networks, POI data, and satellite imagery to classify urban land

use. They used a combination of probabilistic topic models and support vector machines and highlighted the potential for combining remote sensing and social media data for better LULC classification.

Social networks have also been explored for the classification of LULC. Li et al. [10] explored the relationship between spatial technologies and social media in remote sensing, emphasizing the significance of using social media to monitor various aspects of the urban environment. The study highlighted the importance of social media for monitoring natural disasters and land cover. On the other hand, Aubrecht et al. [11] argued that volunteered geographic information (VGI) from sources like Foursquare could enhance real-time LULC classification. However, the accuracy of VGI information depends on the level of participation, which might be low in some regions. Doan and Lim [12] put forth the idea of a “Visitation by Area attraction and Neighborhood” competition to explain the reasons behind people’s check-ins in specific neighborhoods. This study highlights the potential of incorporating information on people’s check-ins in LULC classification models. By understanding this relationship, the authors aim to improve the accuracy of LULC classification. Li et al. [13] combined social media activity with convolutional neural networks (CNNs) to improve remote sensing for emergency response and rainfall event monitoring. The study demonstrated the effectiveness of using social media and CNNs for remote sensing; however, it is limited by the quality and timeliness of the social media data used.

Other studies also utilized deep learning models for LULC classification. Yao et al. [14] proposed a framework that uses POI data and the Word2Vec model to identify urban land use on the scale of traffic analysis zones, offering a unique approach to incorporating social media data in LULC classification. Sithi et al. [15] used geotagged social network images with a Naïve Bayes classifier to explore LULC; however, it is noted that the quality of the geotagged images may limit the performance of this method. Furthermore, Frias-Martinez and Frias-Martinez [16] used self-organizing maps (SOM) to cluster land uses in urban areas based on patterns of tweet activity, offering a novel approach to utilizing social media data. This study provides a unique perspective in the field of LULC classification by exploring the potential of using social media data in this context. However, it is essential to consider some potential limitations of this approach. For instance, the accuracy of the classification results may be limited by the quality and coverage of the tweet data.

Additionally, Weng et al. [17] proposed a deep learning model that utilizes a CNN and a constrained extreme learning machine (CELM), demonstrating promising results in LULC classification. The data used in the study consisted of remote sensing data, including high-resolution satellite images and ground-truth data. The authors tested the proposed model on several land-use types, including residential, commercial, industrial, and open spaces. The authors aimed to improve the accuracy and efficiency of LULC classification by integrating the two models. The study’s results showed that the proposed model outperformed the traditional single model-based approaches, with promising results in classification accuracy. Lastly, Zhai et al. [18] presented a system based on a combination of Place2Vec and POI to identify functional urban regions. This study explores the potential of incorporating POI data into deep-learning models for LULC classification. The use of POI data in combination with Place2Vec allows the system to capture the spatial patterns of land use and the underlying functional relationships between urban regions, demonstrating the integration of POI data into the deep learning model, which enhances the ability of the system to capture the functional connections between urban areas.

In conclusion, LULC classification has been a popular research topic in recent years, focusing on utilizing various data sources such as satellite imagery, social media, and POI data to improve accuracy and efficiency. While many studies have shown promising results, the quality and timeliness of social media data, the quality of geotagged images, and the level of participation in VGI information can limit the performance of some methods. The studies reviewed in this section have demonstrated the potential of integrating social media and POI data into deep learning models for LULC classification, focusing on improving

accuracy and efficiency. Further research is needed to address the limitations of these approaches and to continue exploring the potential of incorporating various data sources for improved LULC classification. However, LUM is a different concept than LULC, and the challenge of identifying LUM is much more complex than LULC.

1.1.3. Land Use Mix (LUM)

Land use mix has been studied in different fields, from transportation to public health and housing market analysis to urban economics. Fundamentally, LUM refers to the mix of varying land-use types present in an area or a building. Having an appropriate level of LUM in an area has many benefits. It can help lower the commuting times and shorten the commute distance [19]. In urban economics, a place with different land types has more potential to raise land values [20]. Additionally, it can encourage people to travel more within an area to discover recreational land-use types such as community centers or parks. These studies demonstrate the importance of LUM in urban areas and the need for effective methods for measuring and characterizing it.

To measure LUM, a number of metrics have been proposed, such as adjacency, intensity, and proximity, to understand the spatial distribution of land use types in an area [7]. Yue et al. [21] used navigational POIs to develop a series of land use mix indicators to characterize neighborhoods along with users' mobile phone activity in 24 h as a proxy of neighborhood vibrancy. Gehrke and Clifton [22] focused on describing temporal aspects of land use mix and guided land use policies and innovative transportation.

Furthermore, Ghosh and Raval [23] presented a study on influence parameters for predicting mixed land use in urban areas. They developed a mathematical model to analyze the relationship between various factors that impact land use patterns, such as zoning regulations, population density, transportation networks, and others. The results showed that this approach could accurately predict the mixture of land uses in a given urban area.

However, to the best of our knowledge, no study focuses on identifying LUM in urban areas in an efficient and automated way using publicly available data sources. This gap in the literature motivated the current study, which proposes an efficient method utilizing POI-based data and Voronoi triangulation to identify LUM in urban areas.

2. Materials And Methods

This section outlines the study area, datasets, and research methodology.

2.1. Study Area

The city of Ankara is located in the center of Turkey (Figure 1). It is the capital of Turkey, with a population of around 5.6 million, and covers an area of 24,521 km². In addition, Kadıköy, a district of Istanbul located on the northern shore of the Sea of Marmara, is included as a second study area (Figure 1). Kadıköy is known for its vibrant urban life and mixed-use development patterns, with a diverse range of residential and non-residential complexes, with a population of 467 thousand and an area of 25 km².

We picked Ankara and Kadıköy as our case study areas. Ankara, being a big city with 24 different districts, demonstrates a significant number of mixed land use cases due to many intertwined residential and non-residential complexes. Kadıköy, with its dynamic urban environment, complements this by providing additional insights into mixed land use scenarios within a densely populated district in Istanbul. In our research, we selected Ankara as the initial study area to develop and confirm our proposed methodology. Subsequently, we utilized Kadıköy as the second study area to validate the final outcomes and compare the results with those obtained from Ankara.



Figure 1. Map of the study areas: (1) Ankara and (2) Kadıköy.

2.2. Datasets

This section will introduce the datasets and preprocessing methods used in our study. The datasets used include publicly available datasets of POI data in addition to land registry data from official sources.

2.2.1. POI Data

There are two primary data sources that we used to acquire POI data for our study. For residential areas, we scraped the data from an ad listings website. For non-residential areas, we used Google Maps API, which provides the locations of non-residential areas. A full list of place types offered by Google Maps API at the time of writing can be seen in Appendix A section. For each request to the API, a maximum of 60 results were returned for a particular area, which is limited due to the nature of Google Maps API.

API calls were continuously made within smaller areas until we identified no additional new POIs. An initial search was performed within an area encompassing a radius of 1500 m ($r = 1500$) without specifying any particular search query, thereby aiming to capture any available POIs. Following this, we systematically reduced the search radius in decrements until reaching 250 m ($r = 250$). For each reduced radius, we repeated the API calls to ensure comprehensive coverage within the smaller segments of our study areas.

This approach enabled us to progressively gather POIs that might have been missed in broader searches. Consequently, by July 2021, we had successfully collected a total of 19,943 non-residential POI locations for Ankara. Similarly, we replicated this methodology for the Kadıköy region in November 2023. By employing the same systematic approach of reducing the search radius and making numerous API calls, we were able to obtain 15,527 non-residential data points for Kadıköy.

For the ad listings data, we undertook a comprehensive web scraping exercise to gather all online ad listings for Ankara from Hepsimlak (<https://www.hepsiimlak.com/>, accessed on 30 July 2019), one of Turkey's largest property ad listing websites, encompassing residential and non-residential properties. Hepsimlak, initially established in 2006, has become a leading platform in the Turkish real estate market. We have conducted data collection from Hepsimlak between June 2019 and July 2019. During this period, we managed to accumulate a total of 84,580 ad listings. Following a rigorous data cleansing process (the details are explained extensively in Section 2.3) we refined this dataset to 60,031 unique ads. The data cleansing involved procedures to remove duplicates, correct inconsistencies, and ensure the accuracy and reliability of the ad listings.

In addition to the data from Hepsimlak, in November 2023, we extended our data collection efforts to include Kadıköy by using the Sahibinden (<https://www.sahibinden.com/>, accessed on 29 November 2023). Sahibinden, established in 2000, is another prominent platform in Turkey that offers comprehensive listings for residential and commercial properties and various other categories and services. Like Hepsimlak, Sahibinden has a significant user base and provides detailed listings.

Using our established methodology, we collected 5942 ad listings from Sahibinden for Kadıköy. After applying the same rigorous data cleansing techniques, this dataset was narrowed down to 4847 unique ads. This additional dataset reinforces the robustness of our study by providing a broader geographical scope and adding temporal depth to our data.

2.2.2. Land Registry Data

The difficulty of reaching the official government data is one of the main reasons we aim to acquire LUM solely using openly and widely available point-based data. Nevertheless, we need ground truth data from the official sources to evaluate the results and compare the predicted LUM values.

Hence, we legally and ethically gathered 650,000 official parcel data in the form of shapefiles from the Turkish land registry website (<https://parselorgu.tkgm.gov.tr>, accessed on 15 May 2020) via various web automation and web scraping techniques on the cloud in a fair usage manner. We used Python's Selenium package to implement the scraper and deployed it on the Google Cloud. We utilized virtual machines with 2 GBs of RAM and 20 GBs of storage on the cloud and managed 100 servers for two months to make the data acquisition faster. We also used Google Cloud Storage to consolidate all the collected data.

The crawled data were packaged as compressed shapefiles containing various information about the land, including but not limited to the borders, area, and land type. These shapefiles were later used to calculate the ground truth results.

One of the vital information included in the official land registry data is the land type labels. After scrutinizing the labels, we found that the following labels can be used to infer the residential areas: home, apartment, housing, domicile, and villa.

Therefore, we labeled the areas associated with the above tags as residential and others as non-residential.

2.3. Data Pre-Processing

We encountered data quality problems with our geospatial data. Many data points appear to be mislocated or mislabeled in our collected dataset.

The descriptive statistics revealed that ads are being published in very close locations. For instance, in an area of approximately 100 m² located tens of real estates, indicating an issue of disguised missing data, which is defined as intentionally and systematically coding missing data as valid data values. In Belen et al. [24], a police station's coordinates were entered as the locations of some traffic accidents in a region, which appeared as a hot spot in the analysis. Similar to this problem, the same coordinates might be systematically used for different real estates.

To solve this problem, we first rounded the latitude and longitude decimals to a certain accuracy after experimenting with a different number of decimal places and considering the GPS signal sensitivity. We have rounded them to up to four decimal places resulting in, on average, a 3.78 m change between the actual locations and the rounded latitude and longitude decimals. Table 1 shows the highest number of POI counts per latitude and longitude pairs rounded up to four decimal places.

Table 1. Number of real estate locations per rounded latitude and longitude.

Latitude	Longitude	Real Estate Count
39.9836	32.8218	201
39.9838	32.8217	153
39.9755	32.8281	112
39.9838	32.8218	107
40.0430	32.8998	97

As a result, we ended up with only 61,096 unique locations for Ankara and 4921 unique points for Kadıköy. However, we still observed that even rounded coordinates of the locations included several ads, which is questionable as it is very unlikely to see so many ads in a very limited area. Therefore, we manually checked these locations. We noticed that several real estate ads are located very close to a real estate agency, indicating that the coordinates of the ads belong to the real estate agency but not the advertised address in the text. For instance, Figure 2 presents a building where tens of real estate locations are advertised as the same building and having a real estate agency in the same building according to the Google Maps data.



Figure 2. Dots represent the real estate located in the same building, and a small image located on the top left presents the real estate agency located in the same spot.

To solve the issue of the disguised missing data problem, we finally used the Google Maps API to check whether there is any real estate agency around each rounded location. We discovered that at least one or more real estate properties are available in places with a high number of real estate advertisements. We kept looking at the vicinity of places with a high number of real estate accumulations until we did not discover any real estate agency nearby. In the end, we found that places with less than three real estate do not have any close real estate agencies within a radius of 250 m. That is why we assumed all the rounded locations having more than two real estates as disguised and discarded. After removing

them, we ended up with 60,031 unique ad listings for Ankara and 4847 unique ad listings for Kadıköy.

We noticed that out of the 650,000 shapefiles in the land registry data, only 45,000 of them were unique. The reason why we have seen lots of duplicate shapefiles result from the small step size we defined in order not to miss any shapefiles. After eliminating the duplicates, we examined data quality issues. We noticed that some of the parcels that we downloaded have issues of conflicting with other parcels. That is, the parcels intersect with others.

Usually, one does not expect to see any intersection with parcels as each parcel represents someone's registered land, according to Turkey's land registry office. Therefore, after carefully examining some of the cases, we noticed that intersections occur in the case of a more general land type associated with the parcel whose areas are much larger than the other intersected parcel. Hence, we discarded the one with the larger area and more general land type. As a result, we ended up with 43,000 shapefiles to further continue our ground truth LUM calculation.

Additionally, to ensure that we have enough data, as mentioned in the previous chapter, we applied the quadrats method and quadrat tests to eliminate fishnets without sufficient land registry data for ground truth LUM calculation. As a result, out of 1862 fishnets, 158 of them were discarded from the evaluation.

We encountered data limitations as well. First of all, we could only calculate LUM for some fishnet eyes. We could not calculate LUM because there is not a sufficient number of POIs in the fishnet to draw a Voronoi. There must be at least four points to draw a Voronoi. After examining the places with less than four POIs, we saw that green areas, wetlands, or agricultural areas mostly cover those areas.

In addition to that, the other reason could be the lack of point-based data, which we collected from an ad listing website for the sale/rent of houses and offices and Google Maps API for various shops, workplaces, and business locations with our best effort.

Lastly, we also saw that some areas do not have an official record of the registry on the land registry website. In Figure 3a, we depict all the available land registry data, but on the left of the figure, it is clear that there is either an apartment or a villa, but the official sources do not confirm.



Figure 3. (a) An urban area covered by all the available official land registry shapefiles. (b) An example of shapefiles having overlaps with each other, indicating a problem with the official data source, overlapping area zoomed within the blue rectangle for better clarity.

Although it is an official data source and each parcel indicates registered land according to the land registry office, we discovered issues of collisions between parcels. In Figure 3b, it can be seen that one big shapefile overlaps with other small shapefiles. In general, we saw that the bigger shapefile usually has a generic label such as land or field. That is why we decided to discard the shapefiles with a bigger area in the case of an intersection.

2.4. Methodology

Figure 4 illustrates the overall workflow of our methodological approach. It starts with the collection of points of interest (POI) data, followed by the creation of a fishnet grid. Within each grid cell, Voronoi are drawn, allowing for the calculation of LUM for each specific area.

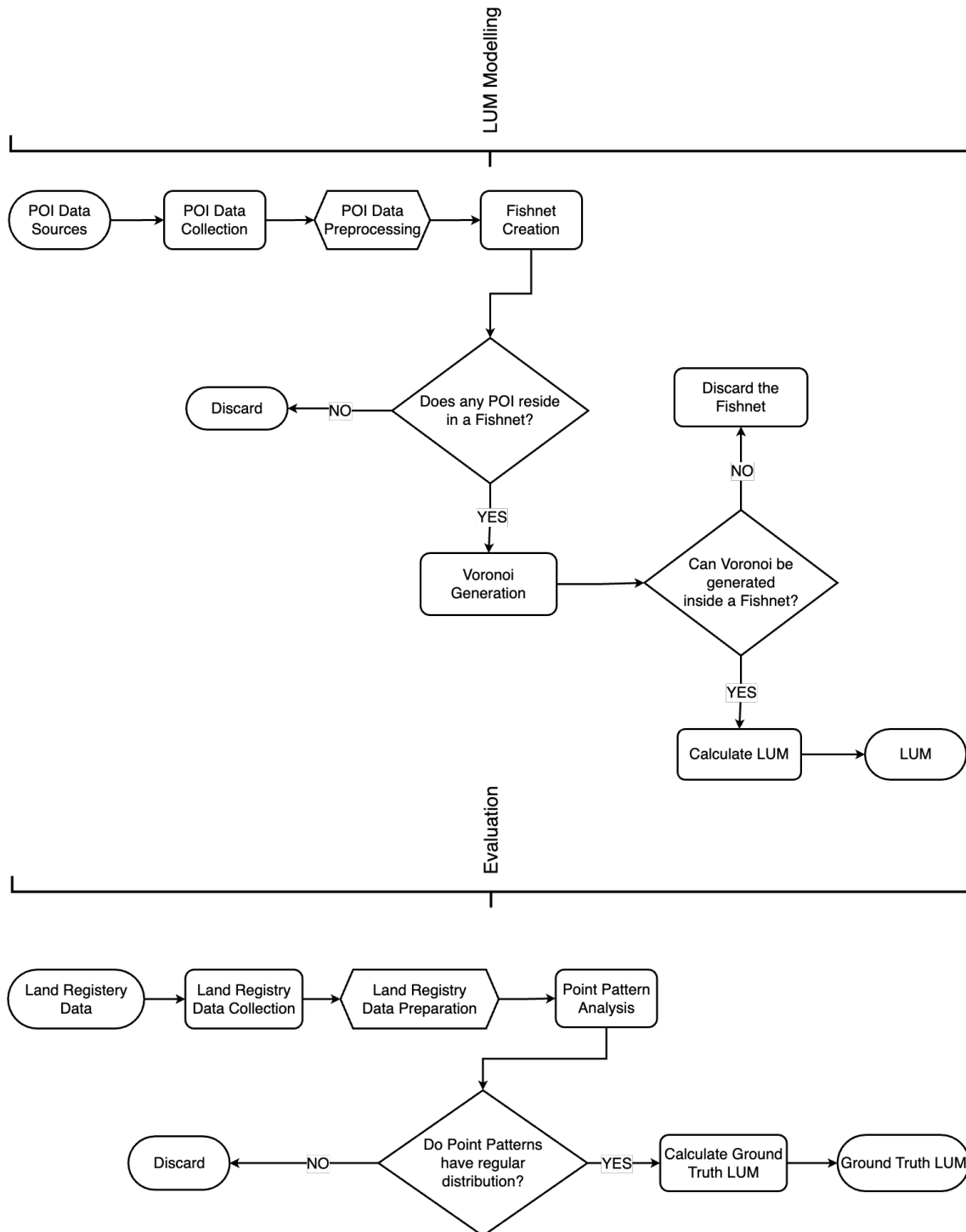


Figure 4. Flowchart of LUM modeling and evaluation of results.

After drawing the Voronoi, we calculate areas of residential and non-residential land use sites depending on the Voronoi area associated with each type of POI, which can be

either residential or non-residential. As a result, we estimate LUM values per fishnet using an entropy-based LUM formula, as explained in detail in Section 2.7.

For evaluation, we collect land use registry data for ground truth from the official sources and apply various data pre-processing techniques to clean the data. Next, we apply a point pattern analysis to understand the dispersion of land registry data. Finally, we calculate the ground truth LUM values and evaluate LUM findings per fishnet with respect to the ground truth results.

2.5. Fishnet Creation

We generate fishnets equally dividing the urban parts of the city using QGIS (<https://www.qgis.org/>, accessed on 30 May 2021), an open-source geographic information analysis tool. Each fishnet eye has dimensions of 500 m by 500 m and an area of around 0.2655 km². POI data gathered from different sources, such as Google Maps API offering non-residential locations and from a real estate listing website showcasing both residential and workplaces, are mapped to each fishnet.

2.6. Voronoi Generation

Voronoi generation is a process that takes a set of points as input and generates a Voronoi mesh. The mesh comprises a set of edges, and each set is shared between two Voronoi cells. The Voronoi cell for a point is the region of space that is closer to that point than to any other.

This step explains Voronoi generation and depicts a generated Voronoi mesh within a fishnet eye. As no area information was present for each associated POI, we created Voronois to estimate an area for each POI, which we will use next to model LUM based on the POI area.

First, we discard the fishnets with less than four POIs as it is impossible to draw Voronois with less than four points.

Using the method of Delaunay triangulation [25], Voronois are drawn, which helps in calculating the possible associated area for each POI based on the Voronoi coverage. Figure 5 shows a generated Voronoi within a fishnet eye with each of the associated POIs.

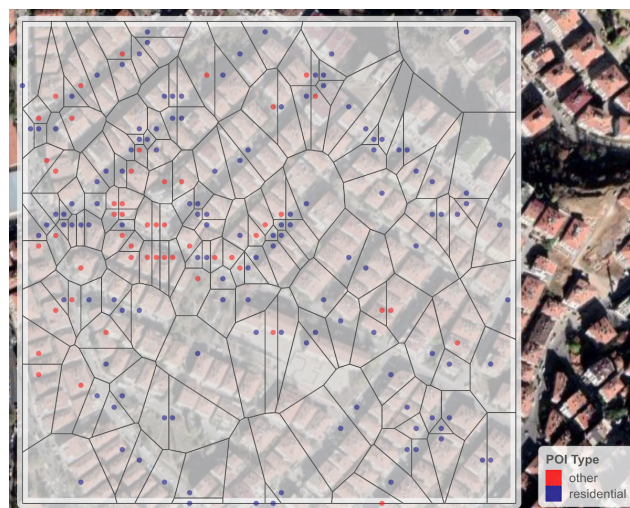


Figure 5. Generated Voronoi based on the POI locations and associated POIs are depicted on top of the Voronoi. Blue dots represent residential POIs, and red dots represent non-residential (other) POIs.

2.7. Calculate LUM

We calculate how much of the area is used for residential or non-residential purposes using an entropy-based method proposed in [5,26]. In the original study, they formulated the LUM index based on the two and six land types. In Equation (1), we extend it to cover n number of land types where a refers to the area of the specific land-use type, k represents

the total area of the particular land use type, and n represents the number of land-use types. Based on the entropy calculation, LUM values range between 0 and 1, where zero indicates no mix, and the higher values indicate a high mix. To calculate LUM, we consider the area of the associated Voronoi of a POI as the area of the particular land use type.

$$LUM = - \sum_i [(a_i/k) \times \ln(a_i/k)] / \ln(n) \quad (1)$$

2.8. Evaluation: Point Pattern Analysis

We analyze the dispersion of surface points covering the land registry area using point pattern analysis to measure the quality of the land registry data collection step. We use a well-known technique, the Quadrats method [27] in our study, which divides a region into equal area sub-regions and counts the number of points in each sub-region to measure the distribution of spatial patterns. As a result, the pattern can be asserted as clustered, random, or regular [28–30] depending on the frequency distribution of the counts compared with the theoretical distribution of spatial randomness [31].

We initially insert regularly distributed synthetic points on each collected land registry area. Our motivation is to understand the adequacy of land registry data via checking the distribution of synthetic points in the area covered by land registry data. If there is a regular distribution of points over the land registry data, it can be concluded as adequate.

An example of surface points can be seen in Figure 6a where yellow shapes represent the land registry parcels, and the dots represent the surface points over parcels. The points are inserted with 0.001 latitude and longitude increments.

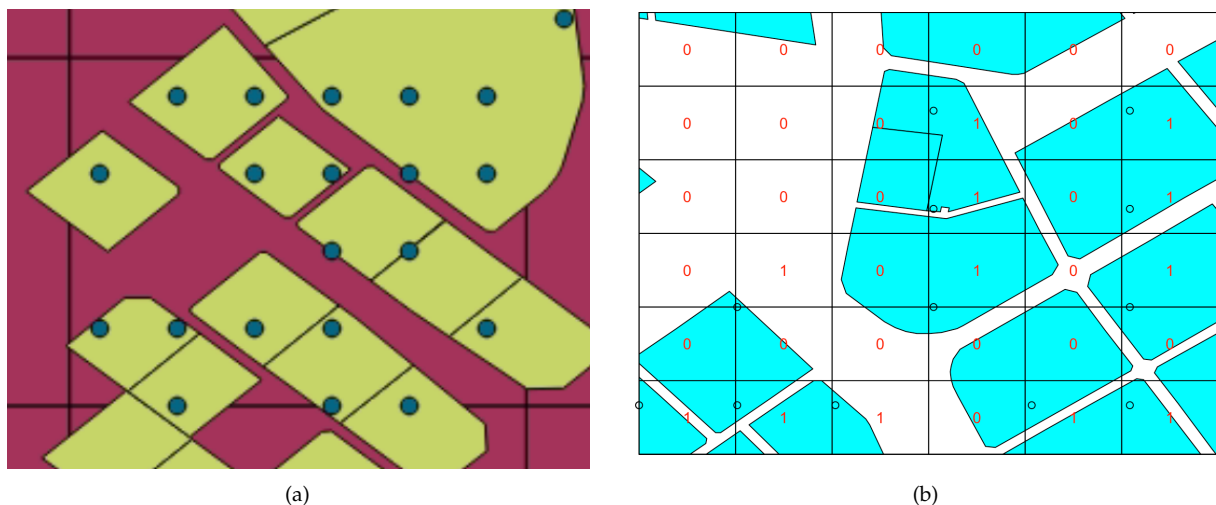


Figure 6. (a) An example of regularly inserted points on each collected land registry data inside a fishnet. (b) Another example of a fishnet depicting quadrants, counts, surface points, and parcels. Small ovals and blue shapes represent the surface points and the intersecting parcels within the fishnet, respectively. The numbers indicate the counts.

To determine the optimal grid size, we tried different grid sizes ranging from 2×2 to 14×14 for the Quadrats method. After that, we calculated the number of regular regions with respect to the total number of regions to select the best grid size for our study. Using a method similar to the elbow method in cluster analysis, we selected the grid size (9×9) as the regularity reaches a certain level afterward. Figure 7 depicts the percentage of regular places with respect to the grid size.

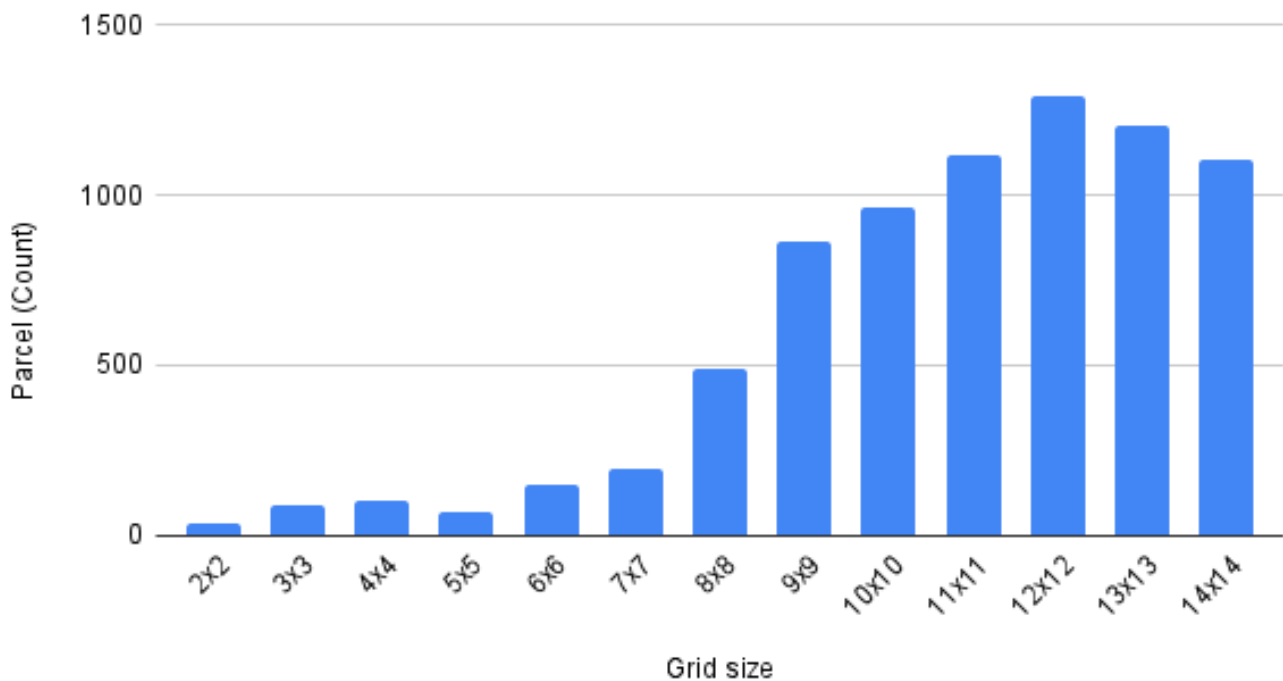


Figure 7. Regularity percentage vs grid size.

As a result, we applied the Quadrats method with 9×9 sub-regions. Figure 6b demonstrates a sample of quadrants and intersected land registry parcels along with the surface points and quadrant counts.

Depending on the quadrat counts, we applied the Quadrat test. As a data preprocessing step, which is detailed in Section 2.3, we only selected the fishnets having a regular surface point pattern with a significance threshold of $p\text{-value} \leq 0.05$.

3. Results

In this section, we first explain how we evaluate our findings and then share the results of the experiments.

Experimental Results

In this section, we unveil our findings and provide the evaluation results concerning different aspects, such as Local Moran's I and the POI intensity per fishnet. Additionally, we present and compare POI-based LUM identification outcomes from these two distinct urban settings.

Using the point-based data from Google Maps API and ad listings data to predict LUM, we present the LUM results on a map to further understand their distribution across Ankara city.

Figure 8 shows the areas in Ankara according to their LUM values, in which gray areas show not applicable results. The colored squares indicate the LUM values of the associated places.

According to the map, as the color becomes dark, the LUM values increase. In general, areas that are far from the city center usually have lower LUM values. On the other hand, the areas with higher LUM values are places that have both residential and non-residential activities in place. Gray areas show the fishnets with not enough POI data, and the darker the color, the higher the LUM.

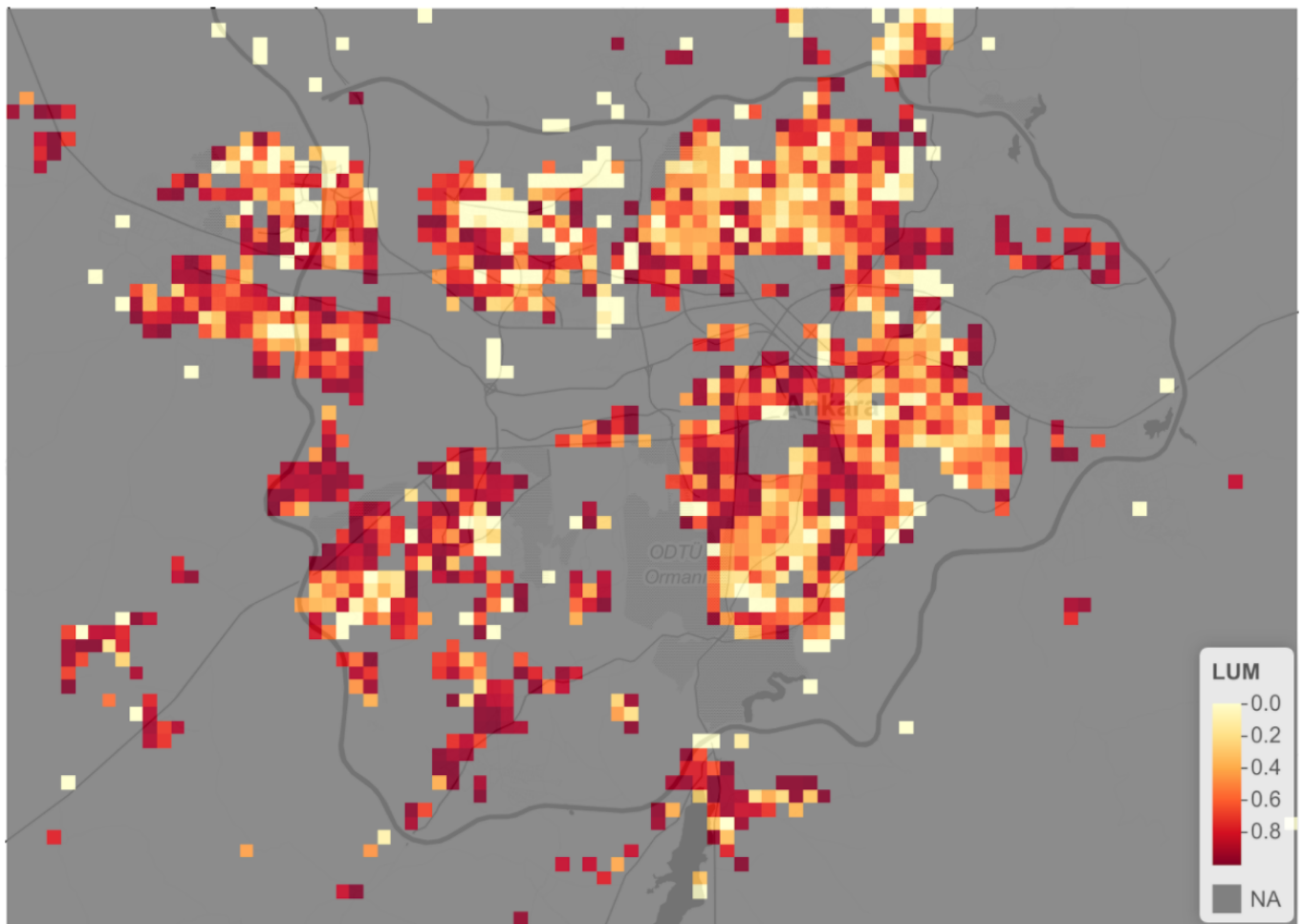


Figure 8. Predicted LUM values are presented on a map. Darker colors indicate higher LUM values.

To evaluate our LUM findings, we first observed the overall results with respect to the ground truth. In general, we wanted to identify whether an area is mixed-use or not. On top of the entropy-based mix estimation, it was essential to label areas as mixed and non-mixed with a confidential threshold. We examined the classification accuracy, precision, recall, and F1-score. In order to classify areas whether they have mixed land use or not, we first defined a cutoff point. For this purpose, we first set 20% of the regions we managed to identify LUM values aside based on a stratified sampling where we obtain samples from places having [0–0.1), [0.1–0.2), [0.2–0.3), ..., [0.9–1] LUM values. We then tested different thresholds on that portion via drawing ROCs for each of the cutoff points, which are depicted in Figure 9.

As a result, we selected the one that has the largest AUC, which is cutoff = 0.5 with AUC = 0.57. After setting the cutoff point to 0.5, we assumed that places having equal or lower than 0.5 LUM values have no mixed land use and others have mixed land use.

Additionally, we also set two baselines to compare the results—one for all the fishnets with no mix representing homogeneous single land use and one for all the fishnets having mixed land use. The overall results are presented in Table 2.

Overall, our only POI-based model's accuracy is 51%, its precision is 48%, its recall is 69%, and its F1-score is 57%. To further understand the results, we scrutinized the areas with respect to neighborhoods and POI density.

In addition to directly comparing the predicted and ground truth LUM values, we also examined the local spatial autocorrelation. Local spatial autocorrelation measures landscape fragmentation in comparison to landscape metrics [32]. It finds spatial dependency of features in space to imply landscape heterogeneity. In this study, we used one of the

local spatial autocorrelation methods, Local Moran’s I, [33] to identify the spatial pattern of different fishnet LUM values.

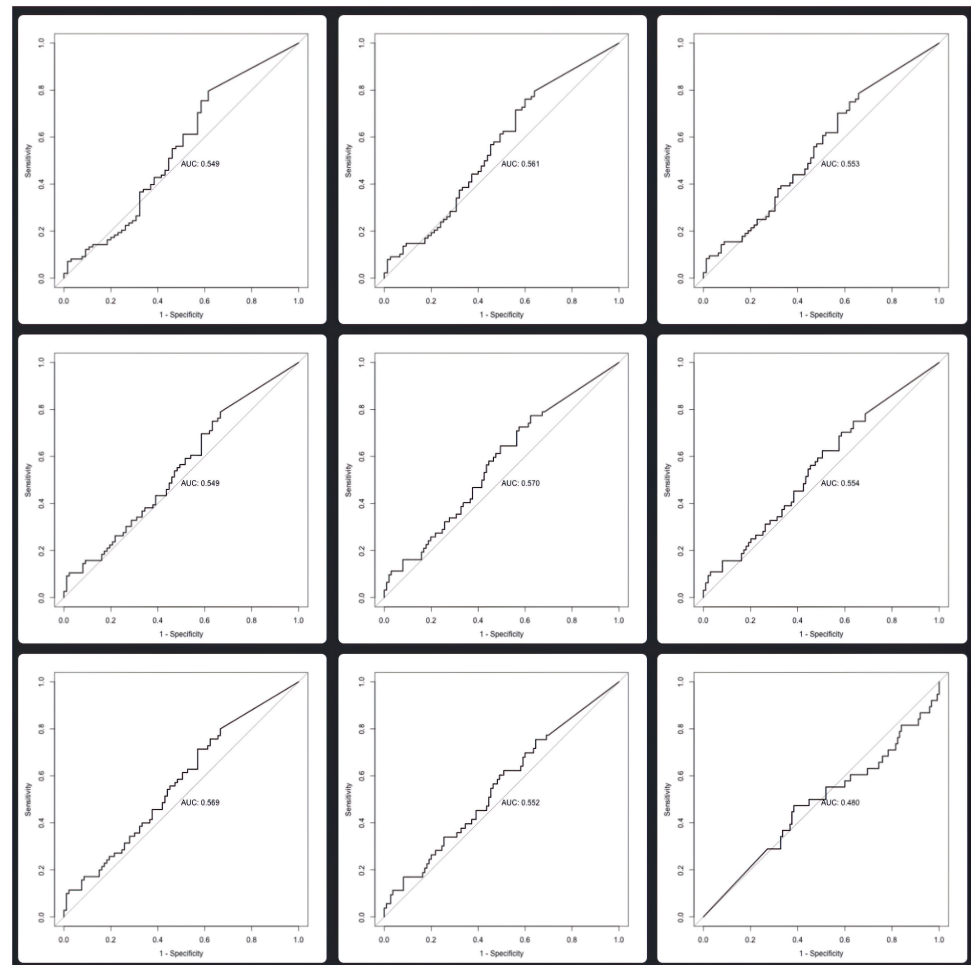


Figure 9. AUC curves for different cutoff points. The top-left AUC plot presents cutoff = 1 and the bottom-right plot presents cutoff = 0.9.

Table 2. Overall accuracy, precision, recall, and F1-score.

	Accuracy	Precision	Recall	F1
Overall	0.51	0.48	0.69	0.57
Baseline all 0 s	0.57	1.00	0.57	0.72
Baseline all 1 s	0.43	0.00	NA	NA

Local Moran’s I fundamentally has two interpretations. On the one hand, it lets you find hot spots. On the other hand, it identifies outliers. Thanks to Local Moran’s I (LMI), we detected local clusters and local spatial outliers to further analyze our LUM classification findings depending on the Local Moran’s I index of each fishnet’s LUM value.

As Tobler’s first law of geography states, spatial distance affects the relationship among things. Hence, ignoring spatial distance among fishnets will make the results biased. Therefore, we calculate the Local Moran’s I for each fishnet based on the assigned LUM values. In Table 3, we presented the classification accuracy, precision, recall, and F1-score for different significance ranges of the Local Moran’s I analysis. The main motivation was to investigate whether the scores increase proportionally to the significance level, and it is confirmed that when the significance of LMI increases, our classification scores also become better. For instance, for *p*-values less than and equal to 0.01, we saw 71% classification

accuracy, 63% precision, 92% recall, and 75% F1-score, with the baseline of all 0 s being 58%.

Table 3. Accuracy, precision, recall, and F1-score for different significance ranges of the Local Moran's I findings along with the number of places residing in each significance level and all 0 s baseline results for comparison.

Sig. Level	Accuracy	Precision	Recall	F1	# of Places	Baseline
$p \leq 0.01$	0.71	0.63	0.92	0.75	28	0.58
$p \leq 0.02$	0.61	0.46	0.92	0.62	38	0.58
$p \leq 0.03$	0.51	0.34	0.92	0.50	49	0.61
$p \leq 0.04$	0.49	0.31	0.92	0.46	55	0.61
$p \leq 0.05$	0.48	0.27	0.92	0.41	65	0.59
$p \leq 0.10$	0.45	0.19	0.92	0.31	96	0.57
$p \leq 0.20$	0.41	0.13	0.92	0.23	135	0.57
$p \leq 0.30$	0.39	0.15	0.70	0.25	161	0.56

Additionally, we look at the classification accuracy, precision, recall, and F1-score concerning the different numbers of POIs residing on fishnets. In order to specify the bins, we first started with equal-depth binning and then tweaked the ranges to make sure that distributions of LUM values in each bin are different from the others. We used a non-parametric Mann–Whitney U test with a p -value ≤ 0.05 to make sure that they were not coming from the same distributions. As a result, we created five bins depending on the POI counts with LUM values coming from different distributions.

In Table 4, the results are presented for different bins such as [4–5], [6–11], [12,17], [18–28], and [29–60]; as a result, we found that the accuracy of finding out whether a fishnet has mix land use or no-mix land use increases up to 65% followed by the 67% precision, 89% recall, and 77% F1-score when the POI count is between 29 and 60. In other words, considering that we are working with fishnets with an area of 0.2655 km², if the POI density ranges between 109 POI/km² and 226 POI/km², we can reach the highest mixed land use classification accuracy of 77% and F1-score of 80%.

Table 4. Accuracy, precision, recall, and F1-score for different POI bins along with the number of places residing in each bin and all 0 s baseline results for comparison.

Bins	Accuracy	Precision	Recall	F1	# of Places	Baseline
[4–5]	0.41	0.34	0.55	0.42	113	0.66
[6–11]	0.42	0.26	0.51	0.35	156	0.54
[12–17]	0.51	0.53	0.55	0.54	107	0.65
[18–28]	0.57	0.56	0.78	0.65	175	0.48
[29–60]	0.77	0.74	0.88	0.80	111	0.56

Finally, we compared the results from Ankara to Kadıköy with the same assumption of having between 29 and 60 points within a fishnet eye, we find the following results presented in Table 5.

Table 5. Accuracy, precision, recall, and F1-score for Ankara and Kadıköy, fishnets having [29–60] POIs, along with the number of places and all 0 s baseline results for comparison.

Bins	Accuracy	Precision	Recall	F1	# of Places	Baseline
Ankara [29–60]	0.77	0.74	0.88	0.80	111	0.56
Kadıköy [29–60]	0.78	0.77	0.91	0.83	18	0.66

Comparing the results from Ankara to those from Kadıköy, we observe that the latter demonstrates a slightly better accuracy and F1-score. This is likely due to Kadıköy's higher density of mixed-use developments which aligned well with our POI-based methodology.

4. Discussion

Dynamically identifying LUM using POI data is a non-trivial and challenging task. This study demonstrates that point-based data, combined with Voronoi triangulation, can effectively identify the LUM index for urban areas. As presented in Table 2, our methodology achieved 51% classification accuracy, 48% precision, 69% recall, and an F1-score of 57%.

Considering local spatial autocorrelation, we observed an increase in accuracy when analyzing fishnets with LUM values similar to their neighboring areas. For fishnets displaying high-high or low-low LUM value clusters, the accuracy reached 71%, accompanied by 63% precision, 92% recall, and an F1-score of 75%, according to Local Moran's I with a significance level of p -value less than or equal to 0.01.

Additionally, an analysis of POI intensity ranges revealed that the number of POIs within a fishnet is critical for accurate LUM value identification. Our results indicated that a higher number of POIs within a fishnet corresponded to increased accuracy. Specifically, when the number of POIs ranged between 29 and 60, the accuracy improved to 78%, with 77% precision, 91% recall, and an F1-score of 83%. This demonstrates the importance of having a sufficient and representative number of POIs to classify LUM accurately.

By comparing the results from Ankara to those from Kadıköy, we observe that the latter demonstrates a slightly better accuracy and F1-score due to its higher density of mixed-use developments. This highlights the capability and adaptability of our POI-based methodology across divergent urban contexts, offering valuable insights for urban planners and policymakers.

These findings emphasize the robustness and applicability of our POI-based approach for LUM identification in diverse urban settings. The results from Kadıköy, in particular, underscored the effectiveness of our methodology in a densely populated urban environment with a high density of mixed-use developments. This further validates that a minimum POI density of approximately 109 POIs per km² is optimal for predicting an area's LUM classification accurately.

Moreover, the method's adaptability allows for real-time updates and dynamic tracking of urban land use changes, essential for responsive urban planning. The high accuracy observed in densely populated areas like Kadıköy underscores the model's potential for application in similar urban settings.

Future research could expand on this by integrating additional data sources, such as social media check-ins or transportation data, to enhance the model's robustness and applicability. Additionally, exploring the method's applicability in different urban contexts can provide insights into its scalability and generalizability.

5. Conclusions

This study presents an effective method for identifying and classifying LUM in urban areas using point-based geospatial data. By employing Voronoi triangulation and an entropy-based LUM formula, we demonstrated that it is possible to dynamically determine LUM without requiring extensive fieldwork or human intervention. The methodology was validated in two distinct urban settings: Ankara and Kadıköy. The results underscore the robustness and adaptability of our approach, achieving up to 78% accuracy and an F1-score of 83% in Kadıköy, where the density of mixed-use developments is higher.

Crucial findings from this study indicate that the number and density of POI within a given area significantly influence the accuracy of LUM classification. Higher accuracy was observed in fishnets with a POI density ranging between 29 and 60, emphasizing the importance of sufficient and representative data for reliable LUM identification.

The comparative analysis revealed that different urban environments could yield varying performance levels. Kadıköy's slightly higher accuracy rates compared to Ankara highlight the method's effectiveness in densely populated areas with complex urban mixes.

Our study offers valuable insights for urban planners and policymakers, supporting efficient resource allocation and the planning of sustainable, mixed-use urban environments. Although the focus was on residential and non-residential land use types, future research could expand this approach to include additional land use categories such as recreational, commercial, and transportation. Additionally, integrating social media data could provide further insights into human interactions with urban spaces, enhancing the understanding and prediction of mixed land use dynamics.

Overall, our study's primary contribution is the development of an efficient, automated method for identifying LUM in urban areas using publicly available point-based data, eliminating the need for fieldwork and human intervention. This methodology represents a significant advancement in urban land use studies, providing a scalable and efficient method for modern city planning.

Author Contributions: Conceptualization, M.A.A.; methodology, M.A.A.; software, M.A.A.; validation, M.A.A.; formal analysis, M.A.A.; investigation, M.A.A.; resources, M.A.A.; data curation, M.A.A.; writing—original draft preparation, M.A.A.; writing—review and editing, M.A.A., T.T.T., S.D. and N.B.; visualization, M.A.A.; supervision, T.T.T., S.D. and N.B.; project administration, M.A.A., T.T.T., S.D. and N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in FigShare at <https://doi.org/10.6084/m9.figshare.25987963.v1>, accessed on 10 June 2024.

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

Appendix A. List of Google Maps Places Types

Table A1. All the available labels in the Google Maps API.

Label	Label	Label
accounting	lawyer	park
airport	library	parking
amusement_park	light_rail_station	pet_store
aquarium	liquor_store	pharmacy
art_gallery	local_government_office	physiotherapist
atm	locksmith	plumber
bakery	lodging	police
bank	meal_delivery	post_office
bar	meal_takeaway	primary_school
beauty_salon	mosque	real_estate_agency
bicycle_store	movie_rental	restaurant
book_store	movie_theater	roofing_contractor
bowling_alley	moving_company	rv_park
bus_station	museum	school
cafe	night_club	secondary_school
campground	painter	shoe_store
car_dealer	electronics_store	shopping_mall
car_rental	embassy	spa
car_repair	fire_station	stadium

Table A1. Cont.

Label	Label	Label
car_wash	florist	storage
casino	funeral_home	store
cemetery	furniture_store	subway_station
church	gas_station	supermarket
city_hall	grocery_or_supermarket	synagogue
clothing_store	gym	taxi_stand
convenience_store	hair_care	tourist_attraction
courthouse	hardware_store	train_station
dentist	hindu_temple	transit_station
department_store	home_goods_store	travel_agency
doctor	hospital	university
drugstore	insurance_agency	veterinary_care
electrician	laundry	zoo

References

- Jacobs, J. *The Death and Life of Great American Cities*; Random House: New York, NY, USA, 1961.
- Swamy, S.; Baidur, D. Managing urban freight transport in an expanding city—Case study of Ahmedabad. *Res. Transp. Bus. Manag.* **2014**, *11*, 5–14. [[CrossRef](#)]
- Jia, P.; Pan, X.; Liu, F.; He, P.; Zhang, W.; Liu, L.; Zou, Y.; Chen, L. Land use mix in the neighbourhood and childhood obesity. *Obes. Rev.* **2021**, *22*, e13098. [[CrossRef](#)] [[PubMed](#)]
- Cervero, R.; Kockelman, K. Travel demand and the 3Ds: Density, diversity, and design. *Transp. Res. Part D Transp. Environ.* **1997**, *2*, 199–219. [[CrossRef](#)]
- Frank, L.D.; Schmid, T.L.; Sallis, J.F.; Chapman, J.; Saelens, B.E. Linking objectively measured physical activity with objectively measured urban form: Findings from SMARTRAQ. *Am. J. Prev. Med.* **2005**, *28*, 117–125. [[CrossRef](#)]
- Xu, Y.; Wang, L.; Fu, C.; Kosmyna, T. A fishnet-constrained land use mix index derived from remotely sensed data. *Ann. GIS* **2017**, *23*, 303–313. [[CrossRef](#)]
- Eom, S.; Suzuki, T.; Lee, M.H. A land-use mix allocation model considering adjacency, intensity, and proximity. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 899–923. [[CrossRef](#)]
- Thanh Noi, P.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors* **2018**, *18*, 18. [[CrossRef](#)]
- Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying urban land use by integrating remote sensing and social media data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [[CrossRef](#)]
- Li, J.; Benediktsson, J.A.; Zhang, B.; Yang, T.; Plaza, A. Spatial technology and social media in remote sensing: A survey. *Proc. IEEE* **2017**, *105*, 1855–1864. [[CrossRef](#)]
- Aubrecht, C.; Ungar, J.; Aubrecht, D.O.; Freire, S.; Steinnocher, K. Mapping Land Use Dynamics Using the Collective Power of the Crowd. In *Earth Observation Open Science and Innovation*; Springer: Cham, Switzerland, 2018; pp. 247–253.
- Doan, T.N.; Lim, E.P. Modeling location-based social network data with area attraction and neighborhood competition. *Data Min. Knowl. Discov.* **2019**, *33*, 58–95. [[CrossRef](#)]
- Li, J.; He, Z.; Plaza, J.; Li, S.; Chen, J.; Wu, H.; Wang, Y.; Liu, Y. Social media: New perspectives to improve remote sensing for emergency response. *Proc. IEEE* **2017**, *105*, 1900–1912. [[CrossRef](#)]
- Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [[CrossRef](#)]
- Sitthi, A.; Nagai, M.; Dailey, M.; Ninsawat, S. Exploring land use and land cover of geotagged social-sensing images using naive bayes classifier. *Sustainability* **2016**, *8*, 921. [[CrossRef](#)]
- Frias-Martinez, V.; Frias-Martinez, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appl. Artif. Intell.* **2014**, *35*, 237–245. [[CrossRef](#)]
- Weng, Q.; Mao, Z.; Lin, J.; Liao, X. Land-use scene classification based on a CNN using a constrained extreme learning machine. *Int. J. Remote Sens.* **2018**, *39*, 6281–6299. [[CrossRef](#)]
- Zhai, W.; Bai, X.; Shi, Y.; Han, Y.; Peng, Z.R.; Gu, C. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Comput. Environ. Urban Syst.* **2019**, *74*, 1–12. [[CrossRef](#)]
- Miller, J.S. Impact of Jobs–Housing Balance on Average Jurisdiction Commuting Times: Virginia Macroscopic Analysis. *Transp. Res. Rec.* **2011**, *2244*, 18–26. [[CrossRef](#)]
- Song, Y.; Knaap, G.J. Measuring the effects of mixed land uses on housing values. *Reg. Sci. Urban Econ.* **2004**, *34*, 663–680. [[CrossRef](#)]

21. Yue, Y.; Zhuang, Y.; Yeh, A.G.; Xie, J.Y.; Ma, C.L.; Li, Q.Q. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 658–675. [[CrossRef](#)]
22. Gehrke, S.R.; Clifton, K.J. Toward a spatial-temporal measure of land-use mix. *J. Transp. Land Use* **2016**, *9*, 171–186. [[CrossRef](#)]
23. Ghosh, P.A.; Raval, P.M. Modelling urban mixed land-use prediction using influence parameters. *GeoScape* **2021**, *15*, 66–78. [[CrossRef](#)]
24. Belen, R.; Temizel, T.T.; Kaygısız, Ö. A data quality case study for Turkish highway accident data sets. In Proceedings of the Road Safety on Four Continents: 15th International Conference, Abu Dhabi, United Arab Emirates, 28–30 March 2010.
25. Field, D.A. Laplacian smoothing and Delaunay triangulations. *Commun. Appl. Numer. Methods* **1988**, *4*, 709–712. [[CrossRef](#)]
26. Frank, L.D.; Sallis, J.F.; Conway, T.L.; Chapman, J.E.; Saelens, B.E.; Bachman, W. Many pathways from land use to health: Associations between neighborhood walkability and active transportation, body mass index, and air quality. *J. Am. Plan. Assoc.* **2006**, *72*, 75–87. [[CrossRef](#)]
27. Bailey, T.C.; Gatrell, A.C. *Interactive Spatial Data Analysis*; Longman Scientific & Technical Essex: London, UK, 1995; Volume 413.
28. Pielou, E. The effect of quadrat size on the estimation of the parameters of Neyman's and Thomas's distributions. *J. Ecol.* **1957**, *45*, 31–47. [[CrossRef](#)]
29. Getis, A. Temporal land-use pattern analysis with the use of nearest neighbor and quadrat methods. *Ann. Assoc. Am. Geogr.* **1964**, *54*, 391–399. [[CrossRef](#)]
30. Rogers, A. Quadrat analysis of urban dispersion: 1. Theoretical techniques. *Environ. Plan. A* **1969**, *1*, 47–80. [[CrossRef](#)]
31. Shu, H.; Pei, T.; Song, C.; Ma, T.; Du, Y.; Fan, Z.; Guo, S. Quantifying the spatial heterogeneity of points. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 1355–1376. [[CrossRef](#)]
32. Fan, C.; Myint, S. A comparison of spatial autocorrelation indices and landscape metrics in measuring urban landscape fragmentation. *Landsc. Urban Plan.* **2014**, *121*, 117–128. [[CrossRef](#)]
33. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.