A Data Driven Approach to Study the Social and Political Statuses of Urban Communities in Kunming[1]

Charles Chang

Abstract

I present a data-driven approach to study the social and political statuses of urban communities in modern Kunming. Such information is lacking in government maps and documents. Using data from a wide variety of sources, many unconventional, I subject them to critical evaluation and computational analysis to extract information that can be used to produce a land use map of sufficient detail and accuracy to allow scholars to address and even answer questions of a socio-political, economic and, indeed, humanistic nature. My method can also be applied to other Chinese cities and to cities elsewhere that lack accurate information.

• Acknowledgements: I thank Zhiwei Ye for his ever-awesome help! Earlier version of this paper is presented at AAG 2014, Stanford, Michigan, and Yale. I thank conference and workshop participants for their useful comments.

Introduction

Digital methods have made rapid progress in humanistic scholarship, including the study of Chinese history. Basically, it is a procedure that calls for translating source, be they textual or cartographic, into data that can be analysed computationally. Data implies number, which can be large. Until a few decades ago, literary scholars and most historians avoided research that depended on the massive use of data; for many a careful reading of a few sources yielded ample insight. Confirming that insight may, however, require data and analysis. Consider the following two hypothetical examples: one from literature, the other from history. Suppose a dispute over the attribution of a Tang poem. Was it by Li Bai, Du Fu, Wang Wei, or an unknown poet? Converting all Tang poetry into machine-readable text (as has already been done) and applying standard text-analysis tools (as is less often done) may give an answer. And if not, the data now searchable, tabulated, and analysed can speak to other questions concerning the poetics of the period. In Chinese history consider the rule of avoidance, in which officials were not to be appointed to their native jurisdiction. How consistent was this practice during the Song dynasty? With the geographic analysis of data on appointments such a question can readily be answered.

Scholars whose research deals with modern and contemporary China face the contradictory dilemmas of insufficiency and over-abundance. Too much data is inadequate, unreliable, or inaccessible. China’s control of politically sensitive data—for example, the number of deaths during the Cultural Revolution or at Tiananmen Square in 1989—is draconian, but control can be strict, if less systematically applied, even in less politically sensitive areas such as the general population’s ethnicity and religious affiliation, or the distribution and economic status of residential types. At the other end is the flood of data from the new information technologies, embraced alike by government, commercial companies, and ordinary citizens. Consider, in image, the trillions of photos made available by street-view surveyors, drones, and smartphones; in words, the immense amount of social media captured on a daily or hourly basis in response to events both trivial and momentous. The abundance and variety of data can overwhelm and confuse, but if selected judiciously with a particular task in mind, they offer a new opportunity to resolve issues which had been set aside for lack of evidence or excessive heterogeneity.

[1]

[1] Scientists have been developing new measurements using “big data.” I list four examples here. (1) Chang, Charles, Zhiwei Ye, Qunying Huang, and Caixia Wang. “An Integrative Method for Mapping Urban Land Use Change Using Geo-Sensor Data.” *In Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, 47-54. ACM, 2015, (2) Hu, Tengyun, Jun Yang, Xuecao Li, and Peng Gong. “Mapping Urban Land Use by Using Landsat Images and Open Social Data.” *Remote Sensing* 2016. 8(2), 151, (3) Zhang, Yuan, Qiangzi Li, Huiping Huang, Wei Wu, Xin Du, and Hongyan Wang. “The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China." *Remote Sensing* 2017. 9, 865, (4) Liu, Xiaoping,

My goal in this article is to demonstrate how we can use a variety of sources and technologies to gauge the social and political statuses of communities in Chinese cities using the emerging metropolis of Kunming as a case study (Figure 1). In its deficiency in statistical information, Kunming resembles many metropolises in contemporary China and, indeed, in developing countries worldwide. Kunming's population, only 147,000 before the Sino-Japanese war, swelled as refugees from occupied north China sought shelter there.[2] After the Communists' occupation of the Chinese mainland in 1949, Kunming returned to its backwater status in remote southwestern China, its growth sporadic and spatially uneven, and this remained the case into the 1980s. From the late 1980s onward the economic boom gave rise to explosive population growth that completely transformed the urban scene. By 2013, Kunming has become a megalopolis of 6.53 million people[3] and yet, for all its size and for all the skyscrapers that make Kunming look modern and efficient, official statistics and maps of communities (or land use maps) are unavailable or grossly uninformative.[4] Official maps, for example, almost never show districts dense with temporary dwellings, yet a stroll even in cosmopolitan Beijing and Shanghai reveals them in all their shabbiness and poverty. Field observation will yield the necessary information, but unless it is conducted by a small army of trained observers, it doesn't give the extent and distribution of urban blight.

Understanding the socio-political scene in China requires an understanding of Chinese cities which in turn requires an understanding of urban communities. In this paper, I use Kunming to demonstrate how a detailed and accurate map of its communities can be produced. The method I use for Kunming is applicable to other Chinese cities and, indeed, to all information-poor cities in the developing world. Such maps provide massive information that was hitherto deficient or unavailable. This article explains how it can be extracted and given easily readable visual form. Into this I inject issues of definition in land use categories that can radically alter the cartographic outcome. My broader ambition is to show that big data has the power, when properly harnessed, to address a wide range of questions raised in social science and the historical disciplines.

[FIGURE 1]

Jialv He, Yao Yao, Jinbao Zhang, Haolin Liang, Huan Wang, and Ye Hong. "Classifying Urban Land Use by Integrating Remote Sensing and Social Media Data." *International Journal of Geographical Information Science* 2017. 31(8), 1675-1696.

[2] Jonathan D Spence, *The Search for Modern China* (WW Norton & Company, 1990), 457

[3] Kunming Statistical Bureau and Kunming team of National Bureau of Statistics 昆明市统计局;国家统计局昆明调查队, Kunming Statistical Yearbook 昆明统计年鉴 (Beijing: China Statistics Press, 2013)

[4] Xi Chen, "State-Generated Data and Contentious Politics in China", *Contemporary Chinese Politics: New Sources, Methods, and Field Strategies*, (Cambridge University Press, 2010), 15–32

Large Data from Unofficial Sources

I now turn to the technical issues in mapping communities in Kunming. I begin with Kunming in the period from 1988 to 2014, when residents witnessed the most change in their urban communities. Mapping them is daunting because archives and documents that trace them to the early years are mostly missing. New and not yet officially designated communities also lack data. The official and unofficial maps I have managed to obtain show well-planned, geometrically arranged boulevards, public squares, and development zones, but I was forced to a very different conclusion when I visited the city and saw slums and communities of highly mixed land use, completely at odds with what the government conceived. I therefore turned to other sources including satellite images from U.S. scientific agency, points of interest (POIs, *xingqu dian*) from commercial mapping companies, geotagged social media posts from smartphone Weibo users who live in Kunming, e-commerce data from online real estate and consumer electronics companies, and street view data from surveying companies. Data from these alternative sources, which are massive, require digitization, computation, and verification.[5] In what follows, I will use a data-empowered framework to incrementally analyse the urban communities by employing computational algorithms capable of extracting the necessary information.

Satellite images come from NASA Landsat. At 30-meter resolution, they can distinguish such categories as built-up areas, forests, and water bodies. Other strengths are: (1) images have had the same spatial resolution since the 1980s; (2) time span coverage is consistent and has a temporal interval of 16 days,[6] short enough to show phases of building construction that is important to my study; and (3) is free and open access. These strengths capture the times of ground breaking and hence the age of the communities. To better obtain information concerning human activities, I collected satellite-generated digital elevation model (DEM) at 30 meters as auxiliary datasets. The topography of these datasets helped me select training samples semi-automatically, a procedure that in turn allows me to infer human activities from the detailed patterns of land use.[7]

[5] By digitization, I mean both converting printed materials, including paper maps, into machine-readable formats and re-structuring information from raw data, for example, converting satellite images to spatial information and geocoded textual addresses into geographical coordinates. For my purpose, computation is using machine learning or other computational algorithms to predict information based on data and verification is ground-truthing the accuracy of the prediction from computation

[6] Due to cloud coverage or technical difficulties such as SLC, sometimes the temporal interval is longer.

[7] Jintao Xu, Tao Ran, Xu Zhigang, and Michael T Bennett, "China's Sloping Land Conversion Program: Does Expansion Equal Success?", *Land Economics*, 86 (2010), 219–44

I also collected the Points of Interest (POI) in 2013. A point of interest is a picture of a part of the city that appeals to visitors. In the old days, peddlers sold paper maps that showed points of interest. The computer age has generated internet maps of far greater appeal and accuracy. Since 2005, mapping and navigation companies, outstandingly Google, Baidu, Tencent, and Gaode have produced information for millions of places (Table 1). Similarly, the volunteer organization OpenStreetMap that supports open data also provides POIs and detailed road networks. These organizations rush to digitize all the POIs as precisely as possible so that they can provide turn-by-turn navigation and other location-based services to consumers. Furthermore, these digitized POIs provide details such as type of business, telephone number, and opening hours, all with a mere click. These kinds of information with such easy availability have never been available before.

[Table 1]

POI data adds specificity to type of urban communities and land use in Kunming, but, although such data can pinpoint the location of an establishment, it cannot delineate its size and boundary. Also, POIs are subject to the state's baneful influence. For example, Google Maps provide fewer POIs in China than in other countries, perhaps because after the company retreated from China in a row over internet censorship in 2010, it has not kept up with POI digitization.[8] Another baneful state influence is the adding of algorithmic offsets on all versions of POIs. I have, however, been able to remove them by using computational algorithms devised by computer engineers.[9] I have also collected all POIs through their respective Application Programming Interfaces (APIs, see details of data collection in Appendix A). POIs from different companies, to the extent that they complement one another, can be used to produce more information. I then check their completeness and preciseness by comparing their geolocations and associated metadata. Compared with Google and OpenStreetMap, Baidu Maps, Tencent Maps, and Gaode Maps are more competitive in that they have been expanding their POIs to provide turn-by-turn navigation and locational based services.[10] Moreover, POIs from these companies

[8] Google, "A new approach to China," published on January 12, 2010, https://googleblog.blogspot.com/2010/01/new-approach-to-china.html, accessed on June 21, 2019

[9] An example of coordinate correction algorithm in Python is retrieved from one of many programs made by geoscientists. I modified it slightly and put it on the Github repository for this article (accessible at https://github.com/placeasmedia/kunming_urban_communities). Only OpenStreetMap is not subject to the geographical offsets that the government requires but it has significantly less POIs than other map providers.

[10] Steven Millward, "Tencent Hits the Road (and Goes off the Beaten Path) to Create China's Biggest 'Street View'", *Tech in Asia*, 2014, https://www.techinasia.com/tencent-maps-covers-china-street-view, accessed on 31 December 2017

have increased the specificity of the information they offer. With the Tencent Map, for example, every POI is labelled at three levels: "Food" at level 1, "Restaurant" at level 2, and "Cantonese cuisine" at level 3.[11] In trying to see what Kunming looks like "from above," I have obtained complete sets of POIs from these three companies. After data collection and pre-processing, I then compare the results thus obtained with hundreds of places that I have validated through field work and close examination of Google Earth images, known for the dependability of their locational information as compared with images derived from other sources.[12] A locational error of less than 30 meters makes my dataset of POIs comparable in accuracy to those derived from Landsat images. I gain further precision by consulting road networks collected from Tencent Maps (after correction) and OpenStreetMap.

POIs are one source of data for constructing the social and political statuses of urban communities in Kunming. Another source are the geotagged posts sent from smartphone social media users throughout the city. Sina Weibo, the Chinese equivalent of Twitter, has smartphone users in the hundreds of millions.[13] I designed a fishnet of points that covers my study area, Kunming, and with it I collected all posts dispatched from November 2013 to July 2014, using Weibo Place Nearby API (Appendix A). The geographical coordinates of posts, as given by the GPS module on smartphones, are highly accurate, with error ranging from 2 to 10 meters.[14] Such a margin may be considered negligible when the resolution of a Landsat image may be in error by 30 meters. Because the geotagged posts can indicate how citizens move at precise time and location, I aggregate them to avoid privacy concerns.[15]

The third source of data is non-spatial, for example, the price of a mobile device as it is listed on Chinese e-commerce websites (e.g. Taobao.com and JD.com, the Chinese equivalent of Amazon.com), price being an indicator of the social media user's socioeconomic standing.[16] Another type of data is information concerning housing projects such as the number of stories of buildings, number of bedrooms and living rooms

[11] Tencent, 2014, "POI classification keywords of Tencent Maps" 腾讯地图 POI 分类关键词, https://lbs.qq.com/webservice_v1/guide-appendix.html, accessed on June 21, 2019

[12] David Potere, "Horizontal Positional Accuracy of Google Earth's High-Resolution Imagery Archive", *Sensors* 8 (2008), 7973–81

[13] According to CNNIC, the number of users is more than 200 million nationwide. Given the percentage of users who use Weibo and the percentage of users who use geotag, the estimated number of users in Kunming should be 200,000. The observed sample is 270,000, close to the estimate.

[14] Paul A Zandbergen and Sean J Barbeau, "Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-Enabled Mobile Phones", *The Journal of Navigation*, 64 (2011), 381–99

[15] The collection and aggregated analysis of Weibo geotagged posts is approved by the IRB at the University of Wisconsin, Madison.

[16] Prices of mobile devices are collected on May 21, 2014

in a sample of units, the selling price and price per square meter of old and new units, be they residential, office, or commercial, as listed on Chinese real estate sale and rental websites on the same date (e.g. Fang.com, the Chinese equivalent of Zillow.com).

Lastly, regarding the visual data, I have also collected street view images in Kunming from Tencent through its API (Appendix A). The street view images provide visual information on exterior appearances that can help distinguish between urban community types, such as urban slums and gated communities. For purposes of comparison and evaluation, I have consulted Google Earth historical images and street views from Tencent Map and georeferenced urban planning maps of Kunming made by the government at various dates.

Steps to identify urban communities

With all the data collected, I map the location, size, and type of communities in Kunming in three steps. Each step incrementally enhances the measurement by integrating information across scales. The first step maps the size, location, and construction year of the land on which the communities are established. To do so, I use satellite images, DEM, e-commerce data, and road network. The second step maps the primary land use categories: residential, manufacturing, commercial, transportation, office, and recreational. No primary category is, however, pure in the sense that it designates exclusively one use. A residential community may have shops at its margins and a commercial area may include residences for its employees. The third step maps urban communities of different socio-political statuses. This step is particularly important for Chinese cities because of its highly mixed urban land use. Here I give two examples. One is the residential category that has been clustered into three hierarchical socioeconomic statuses: urban slum, gated community, and Soviet-style apartment block that was or is a part of a work unit. I choose to map the residential subcategory to illustrate the power of a method, but also, as I shall show later, because residences of different quality reflect the socioeconomic standing of their residents. The other is to separate primary land use category of different political statuses. This includes distinguishing within the office category between commercial offices and the government compounds and, in addition, distinguishing governmental offices within government compounds from the residential areas where many civil servants and tenants live who are unaffiliated with the government.

The three steps I have just sketched correspond with a coordinated effort to extract and integrate information across successive scales. Information at each finer scale offers more detail than the previous one, thanks to information transferred to it. The process is captured in a pictograph in which each applet refers to a step (Figure 2). The rest of this section will give a more comprehensive account of the process.

[Figure 2]

Size, location, and construction year of built-up areas in Kunming

First step. I identify the size and location of communal land, using as my main source more than one hundred Landsat satellite images of Kunming, taken at different times in the past three decades. Satellite images provide spectral information of locations and times that indicate the type of land cover: for instance, if an image shows an area to be green in the summer, I read it as covered with trees or grass. To use a computational method for such identification requires the following procedure: define major classes; map the defined classes using a supervised machine-learning algorithm; and evaluate the result to ensure that the quality of mapping is satisfactory.

The major types of land cover are built-up areas, water, exposed soil, forest, and agricultural land (Table 2). Built-up areas are useful markers in that their annual expansion indicates the size and location of other types of land cover. After economic reform in the 1980s, Kunming expanded rapidly in the next quarter century to more than 70% of today's built-up area. This entire period was covered by satellite imagery. Other types of land cover, rather than being combined and labelled as non-built-up, are included in the classification as separate land cover types. Doing so can improve the accuracy of each type and avoid the possibility of misclassifying areas that have been built on.

I classify the land cover change using a supervised classification method. First, I create training samples that are representative of the entire population of the subjects to be classified. I select the training samples using a semi-automatic method with the assistance of terrain information (e.g., DEM), the latest road network, and seasonal vegetation (e.g., the composite monthly Normalized Difference Vegetation Index). I detail the steps and some results in an appendix (Appendix B).

With all training samples selected, I am able to classify images into land patches of major cover types at a very fine spatial (i.e., 30 meters) and temporal scale (i.e. mostly annually). I then use a supervised machine learning method—Support Vector Machine (SVM)—to classify all the satellite images using the training samples. In comparison with other widely used methods, SVM is superior in its classification accuracy (Appendix B). The classified land forms the basis of land parcels for my other steps, because the farmland developed in the same year in a continuous area is likely to share the same use after urban development. This can be done because in China land cover change is largely in one direction, from non-built-up to built-up.

One remaining challenge is the downtown area, in which change is from built-up areas to built-up areas. I take further steps in inferring circumstantial data: one is to annex single pixels to adjacent continuous pixels in the same urban block if the changes have occurred in recent years. The annexation is based on the rationale that these single pixels register mostly construction extensions of major buildings. Chinese cities, being planned, are unlikely to allow single-family houses the size of single pixels to be constructed separately from such extensions. Another application of this inference is to enlist POIs and the construction years of buildings in the real estate website to infer the approximate year when the renewal occurred.

With the built-up land successfully mapped, I can now lay the foundation for delimiting the size and location of major land use groups and urban communities, within which land cover has changed from non-built-up to built-up in the same year. The method cannot, however, discriminate between adjacent land parcels that have different land uses, the reason being these parcels were built in the same year and were built, moreover, in an urban area that has existed for a long time. To surmount these difficulties, I make use of the current road networks, which are especially dense in the downtown area, for two reasons. One is their relative pattern persistency and the other is that they dissect the urban core area into blocks and further split the land parcels built in the same year along two sides of the road. At the end of all these classificatory procedures, I am ready to consider land parcels, either the small pieces of land developed in the same year or the blocks in the downtown area that have dense road networks, as the basic spatial pattern of land parcels that have the same use.

[Table 2]

Before the land cover map can be used, its accuracy needs be carefully evaluated (Appendix B). Evaluation suggests that my method's first step, which requires digitization of built-up areas on a yearly basis, is successful. Expansion in the peri-urban area, displayed at 30m pixel-size and at one-year or minimal years intervals, accurately shows the size and location of land parcels that are one-time conversions from non-built-up to built-up. The use of phenology, topography, and road network to infer land parcel size and location—a process I have called data-driven—has demonstrably born fruit. Pre- and post-classification processing enables me to capture the change of built-up areas consistently and with an accuracy of annual change achieving 78 percent in the peri-urban area (Appendix B). Cumulatively, by early 2014, the end of my research period, accuracy is above 90 percent.

[FIGURE 3]

To visually compare the results from the traditional approach (i.e. merely relying on satellite images) and the data-driven approach I have just noted in which I rely on a wide variety of data, I selected two typical areas in an urban corridor that stretches from downtown to peri-urban and the countryside. These are: an area at the city centre that has undergone periodic renewal (Figure 3A); and a peri-urban area that has undergone rapid expansion (Figure 3B). My data-driven approach produces consistent results across space, which suggests that it can exploit the feature distinctions across training samples.

Primary categories of land use

Second step. Before I proceed to identify the primary categories of land use, I need to draw attention to how my classification differs from those produced by the Chinese government and leading web mapping companies. I take this step to avoid confusion and to point out certain weaknesses in existing categories. In the past three decades, the Chinese government has periodically updated their definition of land use type in a direction closer to standards set in the U.S. and Europe. Nonetheless, the newer Chinese versions still show government's proclivity to confuse or distort. True, one sees the commonly accepted categories of residential, commercial, manufacturing, public services, and transportation, but government mapping may conflate political agencies with civil services or omit some unwanted categories altogether.[17] Here are two examples. Residences of high-level and low-level civil servants are collapsed into one category, masking stark differences in their prestige and power. Urban villages can be picturesque and a tourist attraction, but they can also be slums damaging to the self-image of political leaders, and for that reason omitted. In contrast to government's proclivity to collapse or omit, web mapping companies boast hundreds of land-use categories. These, however, reflect business needs and consumer demands. Thus, locations of eat places from food stalls to the most luxurious restaurants are listed, as are also shopping amenities that range from old-fashioned stores to the fanciest boutiques. Ever more discriminating definitions of the categories are introduced to keep up with ever growing consumer appetite for variety in goods and services. But, again, urban villages are not acknowledged. They seem an embarrassment to both government and merchant entrepreneurs.

To circumvent these interests and biases of government and business, I define categories by an interest of my own and, so to speak, a bias of my own, which is the political, economic, and social, significance of urban life. For this reason, rather than map specific types of land use, I map primary categories for these most clearly indicate the hierarchical standing of the people who work or live there. That hierarchical standing, revealed in the ways the rich and the poor, the politically important and unimportant, live apart and yet interdigitate is unique to China, for reasons of recent history. After land reform in the 1950s, land was owned collectively. In the 1980s, land and property rights were restored, but even so, new uses of land still showed vestiges of communal use, outstandingly, in the new offices and factories of a designated development zone, but also in residential communities that were rapidly expanding. I outline seven primary categories in Table 2 and Figure 4. They show the political, economic, and social forces at work.

With the primary categories of land use thus defined, how are they to be identified? To address this question, I use the parcels of land and their positions with respect to POIs in the seven categories that are shown in Table 2. Often a land parcel is

[17] The typology from a government copy can be found at "Urban Land Classification and Land Use Planning Standards" [城市用地分类与规划建设用地标准 in Chinese], http://www.mohurd.gov.cn/zqyj/201805/W020180522042539.doc, accessed on June 21, 2019

overlaid with or surrounded by several POIs, each of which carries descriptions of land use in the same three-layered (political, economic, social) structure. It is precisely the political, economic, social forces that shape the urban space and turn the urban development into a variety of communities and entities. To identify the land parcels from the previous step with the correct primary land use category, I make use of the massive amount of POIs that I collect. I code the POIs in a hierarchy of political, economic, and social structure as well as their physical interpretation in assistance of categorizing primary land uses.

To code the POIs, I summarize them by their physical size, spatial relations corresponding to each land parcel, and suggested land use types from their metadata. Based on the size of POIs, I categorize a POI as consisting of building complexes, individual buildings, and sub-size buildings. For instance, a POI of a real estate gated community indicates a large area that contains not only residential buildings, but also other types of building (e.g., bank or shop). By contrast, a POI of a hair salon only indicates a tiny store among other stores on the first floor of a building. In a similar way, a POI that indicates an airport is drastically different from a POI that indicates an ATM. Only a handful of POIs falls under the category of building complex, one that includes factories, real estate communities, and airports, some of which may be acres in extent or occupy an entire city block. Other than building complexes are stand-alone buildings such as museums, libraries, and gymnasiums. Most POIs co-occupy a single building and belong to the sub-building category of store, restaurant, hair salon, and so forth.

I also use the information of distance to POIs in the consideration of land use identification. Rather than using all POIs and giving them equal weight, I designate a threshold of distance to each type and size of the POIs near a land parcel and use the distances as features in the later steps of classification. Specifically, I estimate one feature as to whether a land parcel is located within 500 meters of a POI that indicates a building complex, and another feature as to whether the land parcel is within 150 meters to a POI of an individual building. Moreover, the distance to different types of POIs can be indicative of land use. For example, a parcel of land that is close to shops and real estate POIs is more likely to be residential than commercial. Moreover, I calculate the number of POIs in the types of shops and offices positioned directly on a land parcel. The idea is that the POI of a real estate community surrounded by several restaurants and shops still warrants its being classified as residential; likewise, an office complex that has mostly offices but also a few shops remains an office complex. I include these binary distance features for the POIs of recreation, transportation, manufacturing, office, and miscellaneous types in my classification step (see Appendix C).

Lastly, I extract features from the land use types defined by the commercial mapping companies in the metadata of POIs. As I previously noted, commercial mapping companies all have the proclivity to over-produce POIs for business in pursuit of profit. Nonetheless, these POIs, with their types numbered in hundreds, still can be useful to identify the primary land use categories that I am interested in. I coalesce multiple types of urban land use to simply seven primary categories corresponding to Table 2. To eliminate potential bias of one particular mapping companies, I use both Tencent and Baidu to extract features of land use types. Outstandingly, Tencent Maps provide a broad 20 categories for all its POIs, ranging from healthcare to cultural venues. Baidu Maps, in contrast, offers more POIs of the same establishment at multiple locations to deliver its spatial granularity. For example, a real estate community may have its four gates, on the north, south, east, and west, respectively earmarked on the Baidu Maps. I therefore code the categorical information associated with Tencent Map as a feature and calculate the number of POIs from Baidu Map on each land parcel as other features.

Altogether, I have 23 features defined for each parcel of land: the distance to six sizes of POIs respectively, the distance to eight different types of POIs respectively, the number of POIs in shopping and small business respectively positioned on each parcel, the year of development from the first step, and the size of the parcel. Having all these features prepared, I use a supervised machine learning algorithm, a Logistic Regression Classifier, to categorize the primary use of land according to them. In lieu of the first step, it also takes three steps: identifying training samples, using the machine learning algorithm for classification, and evaluation. I discuss the technical details of the machine learning algorithm in an appendix (Appendix C).

[FIGURE 4]

I evaluate the map's accuracy by using the testing sample. The results of evaluation show that land use classification reaches an acceptable level of accuracy. Overall, it reaches 78 percent. Variation from one category to the other is, however, large. Thus, in commercial, office, recreational, and residential categories, accuracy reaches levels of 88 percent, 60 percent, 67 percent, and 80 percent respectively. This is satisfactory, but in manufacturing and transportation, it is lower.

To help inspect the evaluation's results visually, I make a map that covers both the downtown and peri-urban areas of Kunming in major land use categories (Figure 5). It is quickly obvious that the downtown area is primarily for residential use with, however, other uses such as upscale commerce, office, and recreation mixed in. The area reserved for recreation includes buildings of historical significance, for the city's draw goes beyond mere housing. Outward from downtown, land uses are mixed and include residences and industries that have had to move out when the urban core expanded, newly-built residences for new residents, new commerce and industries, as well as old stores and factories that have somehow escaped removal. Still farther out from the city core is the peri-urban area where land uses are even more mixed and include villages, isolated farmsteads, new communities, offices and industries that seek not only cheaper land but also escape from the governmental regulations that apply to the urban core. Residential use being dominant in the city, the relative value of its subdivisions provides an index to the socio-economic status of the people who live there.

[Figure 5]

Urban communities with different socio-political statuses

Third step. I turn to residential land use's subcategory, which is made up of gated communities, work-unit communities, and urban slums. Gated communities are mostly walled real estates, inside which are either high-rise buildings or small condo apartments, surrounded by grassy lawns or landscaped gardens.[18] Architecturally distinctive, gated communities symbolize high social status to the Chinese nouveau riche. Work-unit communities are apartment buildings of six to eight stories, built in the Soviet-era style and compactly arranged.[19] They once catered to workers who belonged to work units, but now are largely transformed into tenement housing for people of modest income. Urban slums are small houses of one to three stories, confined in small urban spaces, and of substandard construction. Some of them were villages surrounded by farmland but became engulfed by new urban development. Others were built in the process of urbanization, as the demand for affordable housing surged and the supply for it was lacking. Urban slums frequently turn into havens for migrant workers, students with a degree and little else, and the unemployed—in short, a ragtag assortment of people trying to make ends meet.

Where and how people live matter, yet this topic is underrepresented in works by scholars of modern China. The reasons? Residential communities are poorly documented; including internet maps. Not all urban communities have POIs. Even if they do, their size and location remain uncertain because their POIs are not indicative of their size and boundaries. In some work-unit communities and in almost all urban slums, there are no POIs to draw attention to. For example, dormitories in schools, hospitals, and state-own enterprises are often left out. POIs exist for these institutions, but not necessarily for their dormitories. In government compounds, there are residences that were once exclusively used for government officials but are now open for sale or rental to others. These, too, cannot be taken as government offices. In the same manner, the residential part of small restaurants, convenience stores, and hair salons in urban villages is edged out by a multitude of commercial uses and are often neglected by the government statistics.

I first classify residential use into gated communities and work-unit communities, based on their physical characteristics as well as their price and the availability of services. To locate and map gated communities, I use detailed information from a number

[18] Zhang, Li, *In Search of Paradise: Middle-Class Living in a Chinese Metropolis* (Cornell University Press, 2012), 109-115

[19] Some scholars address this kind of community as communal apartment block. In this article, I use it to indicate the architectural style of apartment buildings, instead of a social institution. The architecture for all the communal apartments in a block is often similar and utilitarian. Individual household often occupies an apartment while multiple households share the corridors or some common spaces.

of sources, including (1) the sale and rental record on a real estate website, which supplies information on basic construction type, whether the buildings be villas or high-rise (>10 stories), price of an apartment, number of bedrooms and living rooms in a unit; (2) POIs that provide supplementary information which, by indicating the type and number of surrounding services, can change a community's classification from gated residential to work-unit residential, or to some other category altogether. The information of the gated communities and work-unit communities listed on the real estate website, including the addresses and size of the urban community, often have been verified by the commercial companies and are thus accurate. On the other hand, the number of POIs that indicates shops and small businesses on the land parcel can identify the work-unit communities, for these communities often have a large number of shops and services at the ground floor. This helps identify gated communities and work-unit communities. Dormitories in the government compounds are also identified.

The information about urban slums are less accurate for the reason noted earlier: documents about these communities even on commercial websites are quite often incomplete and lacking. For this reason, I incorporate social media data and obtain from it the temporal pattern for each land parcel. The temporal pattern of social media posting at residences may confidently be assumed to differ from the temporal pattern of workplace in three ways (Appendix D). First, users of social media, while at home, are likely to send more social media posts during off-work hours, early in the morning and late at night than during work hours on weekdays. Second, they are likely to send more from 9 am to 5 pm during weekends than during weekdays. Third, they are likely to send more during holidays and special occasions when they stay at home for family gatherings (e.g., the Eve of Spring Festival). Based on these criteria, I designate land parcels that were previously categorized as commercial as urban slums to the extent that the temporal pattern of posts satisfies two of the three ways noted above.

[Figure 6]

To evaluate the classification result for the sub-category of residential use, I randomly select a sample of points with the assistance of Tencent Street View and Google Earth high resolution images: 56 points in gated communities, 57 in work-unit communities, 10 in government residence, and 62 in urban villages. I then assess the accuracy of the computational classification by comparing them with the correct labels. The result shows my method to be accurate. Gated communities are classified at an accuracy of 84 percent, work-unit communities are classified at an accuracy of 42 percent, government residence are classified at an accuracy of 100 percent, and urban slums are classified at an accuracy of 74 percent.

A visual inspection of Figure 6 shows how these communities are distributed. In the downtown area where many POIs indicate land use and where social media are abundant, gated and work-unit communities are easily distinguishable from one another, even though both occupy already developed space. In the peri-urban area, where built-up patches are interrupted by natural features such as lakes and green space, gated and work-unit communities are even more clearly bounded and so distinguishable. Urban slums, by contrast, are fragmented and shapeless. They do not fit into zones and appear to scavenge for urban spaces that are either undesirable or taken up by other uses.

Discussion

My map of communities across time and space shows that detailed information concerning the political, economic, and social nature of Chinese cities can be obtained by means of a data-empowered method, when little information comes from official sources. My three-step approach is complemented and enhanced by using not only datasets, both new and newly accessible, but also by knowledge acquired from various disciplines and, importantly, a willingness to interpret, based on my extensive exposure to statistics. What have I achieved? What is the potential of my method? The core content of the paper should answer the question, but I would like to add certain findings that I could not emphasize in the body of the text without stalling its narrative flow. These are: urban expansion and renewal, mixed-use spaces, and a data-driven approach.

Urban expansion and renewal

Urban expansion, particularly in the United States, is usually understood as urban sprawl, or to put in another way, occurring along a horizontal plane. As such, it is accurately recorded by satellite images, which offer the view from above. However, in Kunming, the city has also expanded vertically, as any side-view of skyscrapers makes clear. Horizontal expansion shows a clear pattern of government stimulation and control, for most of it follows along or near the road networks that the Kunming municipal government has built and away from wetland near Dianchi Lake and forest areas surrounding the city. Thus, to a degree, government urban planning has been effective. As to expansion in the vertical direction, urban change in recent decades is from an expanse of scattered low buildings, with a clutch of taller ones at the centre, to a forest of sky-reaching giants. Kunming urban planners apparently have no defined skyline to guide vertical growth or even care about their city's skyline as American planners often do theirs. I reconstruct the sideview from the resident's use of smartphone and their real estate trading records online. From these two views, side and vertical, I gather information of, say, gated communities in Kunming, their social standing, the services they require, and the newly paved roads circulating in their midst (Figure 3. B2).

Mixed-use spaces

My maps of land use in primary categories and of urban communities also show highly mixed use. In Kunming, the government of Kunming municipality is reallocated in the newly designated development zone, while the provincial government is located in the old city proper. In the newly designated development zone, land use is planned and often takes a geometric form. In this zone, government buildings are surrounded by gated communities where high officials and the rising class of businessmen live. Mixed within this zone are commercial establishments that cater to the needs of gated-community residents (Figure 5). In the old city proper, the provincial government buildings are surrounded by business offices, gated communities, work-unit communities, and recreational spaces.

As for urban slums, ramshackle houses are a common sight (Figure 6). They do not, however, constitute an expanse of poverty and urban decay that one often finds in US cities. Instead, poor and struggling people in Kunming live in fragmented patches of land, as the term "urban village" or the more universal term "urban slum" suggests. Moreover, gated communities, work-unit communities, and urban slums intermingle to a surprising degree, apparently in disregard to differences in social status. Also surprising, because in apparent disregard of the prestige of office, is that government residences, too, are scattered across the city, some in gated and some in work-unit compounds. In this indifference to the use of residential zoning to reflect socioeconomic status, Chinese cities are more like old European cities than are cities in America where spatial location unambiguously reveals one's social standing. Fieldwork done by a team of urbanists will gain a view of Kunming that shows details of land use, their political and economic implications, of the sort I have sketched above. My paper makes the case that, in the absence of such a team—in the absence of even the possibility of such a team producing a work that is then made available to the public—a realistic map of the city can still be constructed.

A Data-Driven Approach

In the study of Chinese cities, crucial empirical evidence such as fine-grained land use, residents' social network, their social or political preference, is simply missing. Even if evidence exists, it may come from a small number of observations at snapshots of time and place and thus lack representativeness for the entire community or city. The data-driven method I propose solves the representativeness problem by collecting data from as many sources as necessary and then integrates the data collected with the method's two principal components, the inter-scalar and the circumstantial. The former is integration at different scales.[20] The zooming camera provides a simple example. From the air, the camera's lens opens to cover an entire neighbourhood, then narrows until it targets a house, in the process of which more information is gained about the house than could have occurred without the prior view of the entire neighbourhood. Zooming enables information from what one sees at one scale to be transferred to and integrated with what one sees at another scale.

Circumstantial integration, the other component I propose, infers from indirect evidence that may capture the nature of a place, a person, or an event for which no direct evidence is available. Such evidence needs to be abundant and digitizable for firm conclusion to be drawn. At the opposite extreme in the use of circumstantial evidence is

[20] Note that computer scientists often refer to scalability as a modular design by distributing the computation across small scalable components concurrently, which is an entirely different concept from what I propose.

to draw a conclusion based on a single fact and a negative one at that. Was China in the ninth century a peaceful and well-governed empire? Can one say "yes" because the Japanese monk Ennin, who travelled widely over it, did not record in his diary a single encounter with brigands?[21] More convincing as indirect evidence is where numerate data is available, but most convincing of all is where abundant digital data exists. A case in point is the socioeconomic standing of residential compounds in a modern Chinese city. I do not have direct evidence of their standing, but I do have indirect evidence, namely, the movement of bureaucrats in and out of government buildings, their times of arrival and departure, and more importantly the regularity of these times, for irregularity is an indication of the bureaucrats' high status and their residential compound's prestige.

Conclusion

Digitization has significantly expanded the scope of the social sciences and humanities and will no doubt do even more for them in the future. Certain areas remain, however, unaffected, the most obvious of which is value. In literature, digitization may help one see that Shakespeare's vocabulary is far larger than that of his French near contemporary, Racine (24,000 to 2,000, according to Arnott[22]), but that fact does not allow one to judge as to who is the superior playwright. In the social sciences, digitization has much to say about how institutions work, but little about what constitutes a good institution. Besides value judgment and conceptualization framework, another area that seems to lie beyond the need for digitization is culture. As ethnographers understand the term, culture consists of the distinctive customs and practices of small, isolated human groups. Understanding their way of life or culture calls for prolonged field exposure, empathetic observation and interpretation, rather than skill in computational methods.

China's ethnic cultures engage the attention of ethnographers. China itself is, however, a large modern society, the study of which is aided by the sort of digitization and computation I have explored. Modern society is "rational" to the extent that it is based on a structure of power, structure being geometric and abstract as compared with, say, power's sartorial insignia, the Chinese emperor's brocaded yellow robe, for example. Modern society is also rational to the extent that its economic base depends on a marketing system that operates impersonally rather than on the passions and desires of select individuals. To study modern Chinese society, I have therefore used a data-driven approach that is itself, in intent, rational and abstract.

Critical thinking is at the core of literary and humanistic study. I hope I have also demonstrated it in my use of the data-driven approach. Deciding on what is and is not appropriate data, its possible biases, the right algorithms in analysis, and so on, is critical

[21] Edwin Oldfather Reischauer, *Ennin's Travels in T'ang China* (New York: Ronald P, 1955), 138

[22] Arnott, Peter D, *An Introduction to the French Theatre* (Rowman & Littlefield Pub Inc, 1977)

thinking that the data-driven scholar must command. But such a scholar has an even closer affinity with his humanist counterpart, which is sensitivity in the use of language—the meaning of words. To study Chinese cities, one cannot simply take over, for example, the land use categories of American planners, or use the words "suburb" and "slum" of Chinese cities without qualification when they invoke quite the wrong images.

Throughout this paper I have contraposed, at least implicitly, between an Olympian's perspective from high above ground and a field worker's perspective on the ground, between one who generalizes and one who particularizes, between the theorist and the empiricist. Interestingly, "theory" is Greek for "speculation" or "spectator."[23] The ancient Greeks seemed to have favoured the spectator who, high on the bleachers of the stadium, sees the players at their game and understands what is going on. The players themselves, covered in dust and sweat, are too engaged to know. So the spectator (theorist) thinks. Naturally, the players think differently: they who are aware with their body's every straining muscle where to run or throw the ball are the ones who truly know the game. Both spectator and players have their point. In this paper, I have taken more the spectator's position in that I have tried to show how a data-driven approach can contribute to not only seeing the pattern of life, but also—here and there—its dense details.

[23] Nightingale, Andrea W. "The Philosopher at the Festival: Plato's Transformation of Traditional *Theōria*," in *Pilgrimage in Graeco-Roman and Early Christian Antiquity: Seeing the Gods,* edited by Jaś Elsner and Ian Rutherford (OUP Oxford 2007): 151-180.

Table 1. Data Sources

	Satellite Imagery	POIs[1,2]					Road Network[2]	Social Media	Real Estate Record[3]	E-commerce Record[4]
Sources	NASA Landsat	Tencent	Google	Gaode	Baidu	OpenStreet Map	OpenStreet Map, Google Maps, Tencent Map	Sina Weibo	Fang.com	Taobao.com, JD.com
Total number of records	71	192,468	70,428	472,290	101,584	2,049	14,972 km	1,481,478	1,809 (1,393 transactions in 397 communities)	145
Geographical error (meter)	0	5	394	20	3	11	<5	<10	~100	N/A
Metadata	N/A	Name, address, land use type	Name, land use type	Name, address, land use type	Name, address, land use type, consumers' review	Land use type	N/A	Post id, user id, sent time, sent device[4], content, user's information such as home province and city	Community name, address, price, built year, building type, total building size, total land size, total units	Smartphone brand, type, price

[1] Geographical coordinates are corrected from GCJ-02 to WGS-84. POIs by each company are evaluated with 10 random sample respectively.

[2] Geographical coordinates are corrected from BD-09 to GCJ-02, then further corrected to WGS-84, which ends up with an error that is less than 5 meters due to this correction.

[3] All real estate records are geocoded according to their addresses, which leads to a small geographical error that is about 100 meters. Consider the size of the real estate communities, the error is negligible.

[4] E-commerce record of mobile devices of major brands are collected in May 2014 to match those used in Weibo. Some hundreds of small brands are generalized as phones under 1,000 RMB.

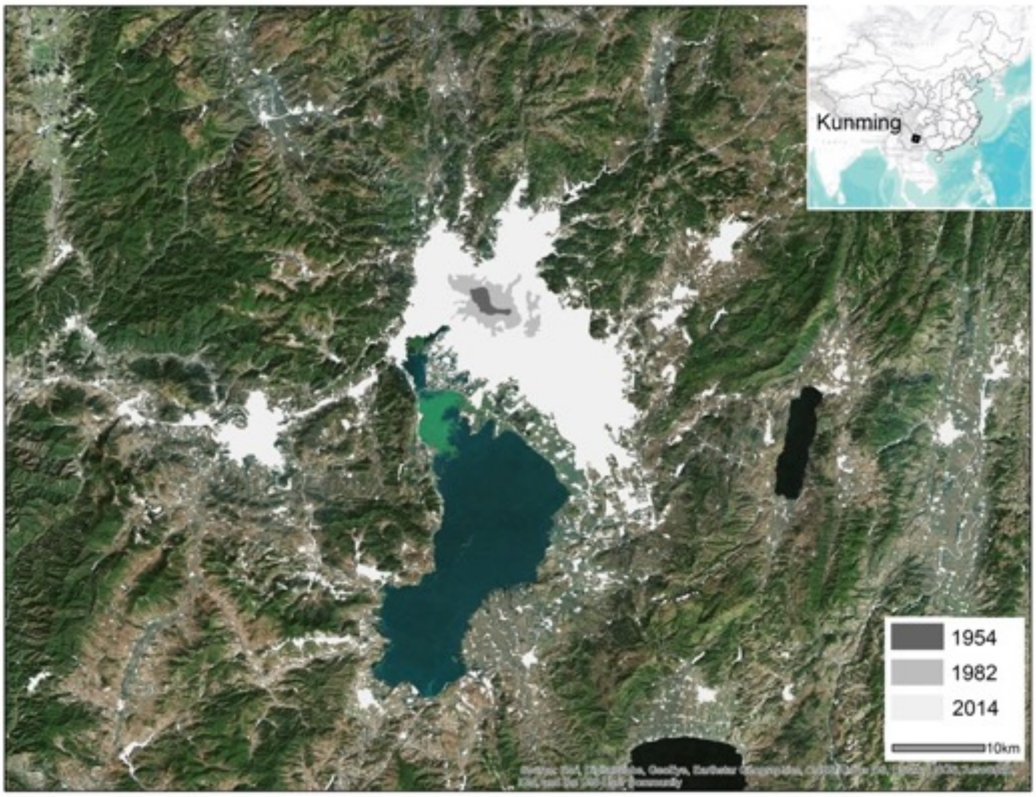
Table 2. Major types of land covers, primary land use categories and communities

Land Cover Types	Primary Categories of Land Use	Subcategory Communities	Detailed Land Uses
Built-up areas[1]			
	Manufacturing		Factories, workshops
	Transportation		Airport, railway station, toll stations, river dam, road appurtenances, etc.
	Office		Corporations, companies, banks, government offices, office buildings in schools and hospitals, etc.
	Commercial	Shopping center	Large shopping malls and shops, restaurants, night markets, KTVs, car shops, fitness clubs, and other individual buildings that are designated for shopping and services.
		Urban slums	1-3 stories small buildings where residences are located at the upper level or backside of the building and shops and eat stall are at the street level and front side of the building.
	Recreational		Convention center, museum, library, buildings in park and zoo, etc.
	Miscellaneous		Farmhouse, graveyards, storage, unused land, etc.
	Residential		
		Gated communities	High-rise apartment building complexes, villas, or a large scale of apartment buildings that are often gated and guarded, with few exclusive living services, such as spa or veterinary center, suited
		Work-unit communities	6-8 stories Soviet Union style apartment buildings that was once built by work units such as state-owned enterprise, but now is accessible to the public, with the lower levels as stores and restaurants
		Government residences	Similar to the work-unit communities except that they are located in a government compound that is sometimes walled or guarded.
Natural elements[2]			
Forest	Urban trees		Street trees, urban park forest, etc.
Grassland	Urban garden		Lawn, garden, urban farm, etc.
Water	Water		Lake, manmade pond, open sewage, etc.
Exposed Soil	Undefined or unused land		Construction site, waste landfill, mine pit, unused land, etc.

[1] Built-up areas here refer to the land surface that is manufactured and can be detected using remote sensing images, regardless of specific human use. In contrast, major land use categories refer to the human use of urban land, roughly corresponding to the major categories of zoning ordinance and urban planning practices.

[2] Natural elements co-exist as various detailed land uses in urban environment. However, I do not discuss them in detail in this research.

Figure 1. Study Area



This map shows the expansion of Kunming over time. The digitized urban area circa 1954 came from unclassified US military maps (U.S. Army Service 1954); the one circa 1982 came from Chinese government (1982); and the one circa 2014 is a part of my result.

Figure 2. Workflow

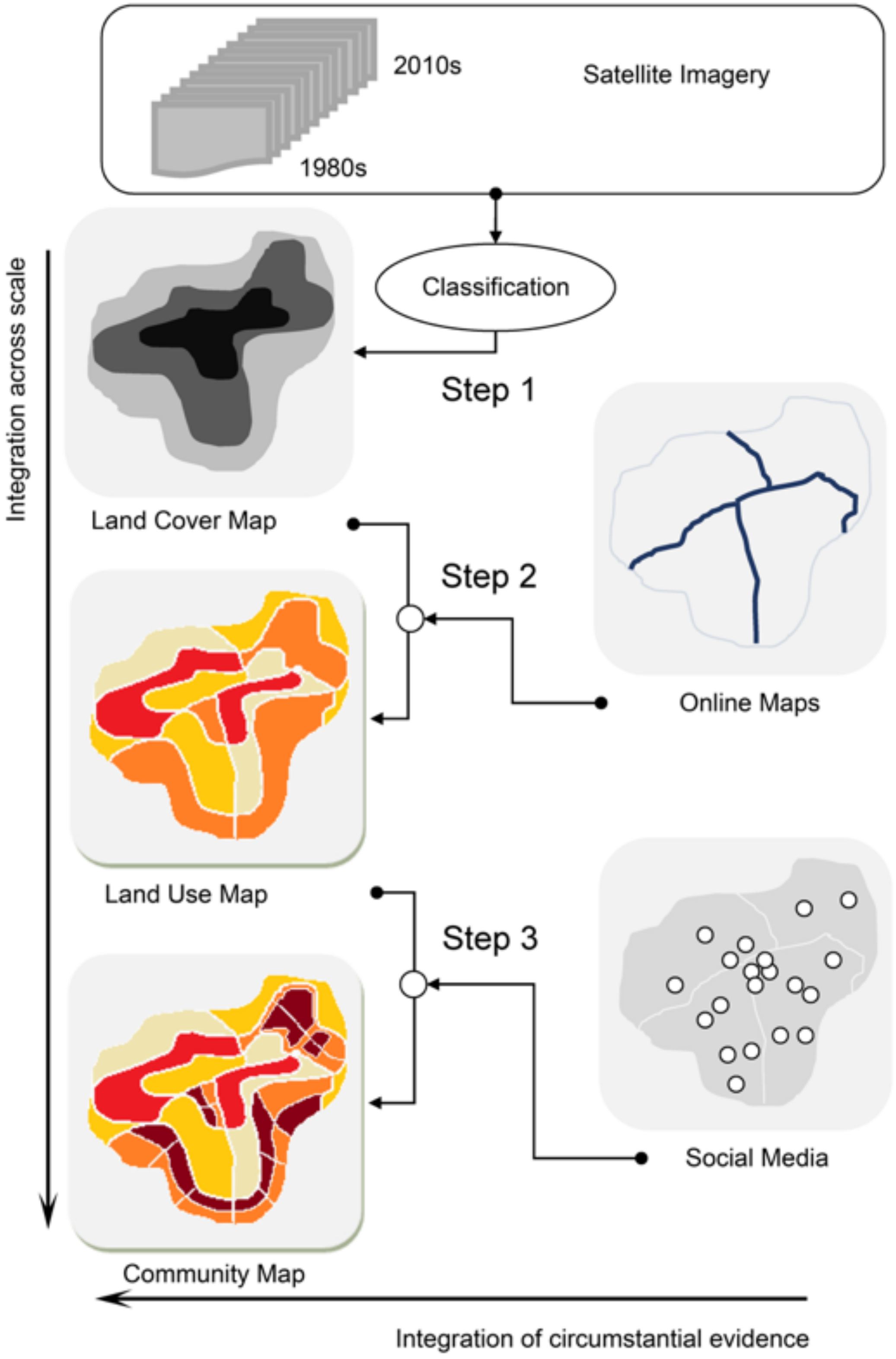


Figure 3. Land Cover Change in Downtown and Peri-Urban Areas

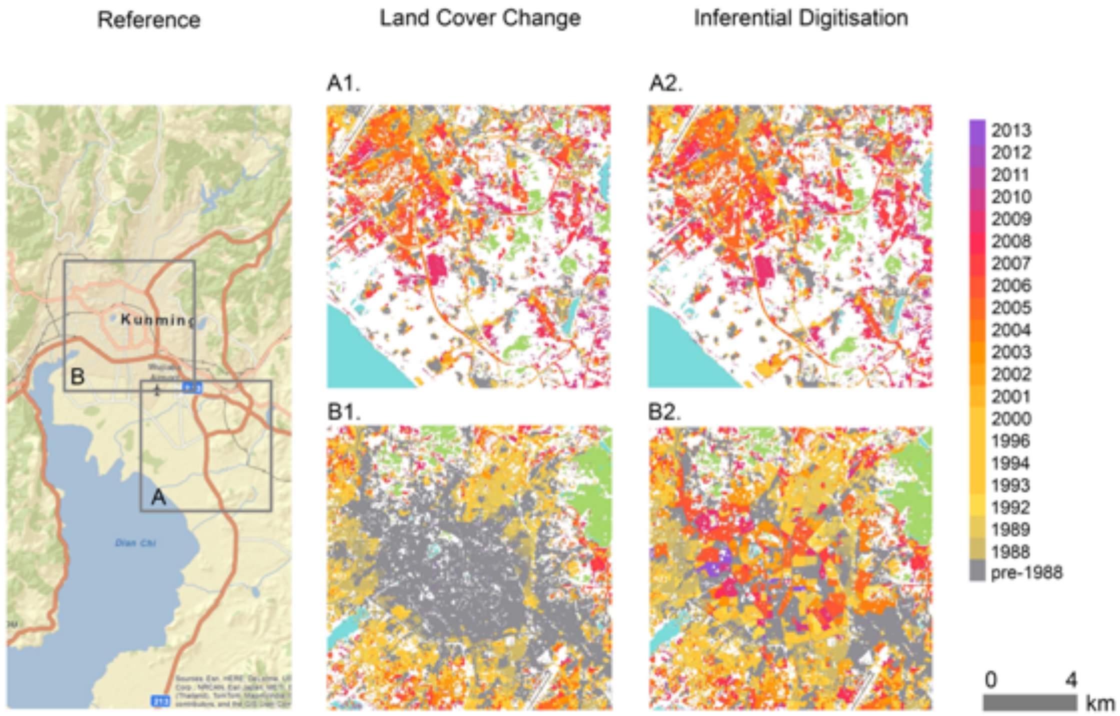


Figure 4. Layout and views of typical land uses and communities

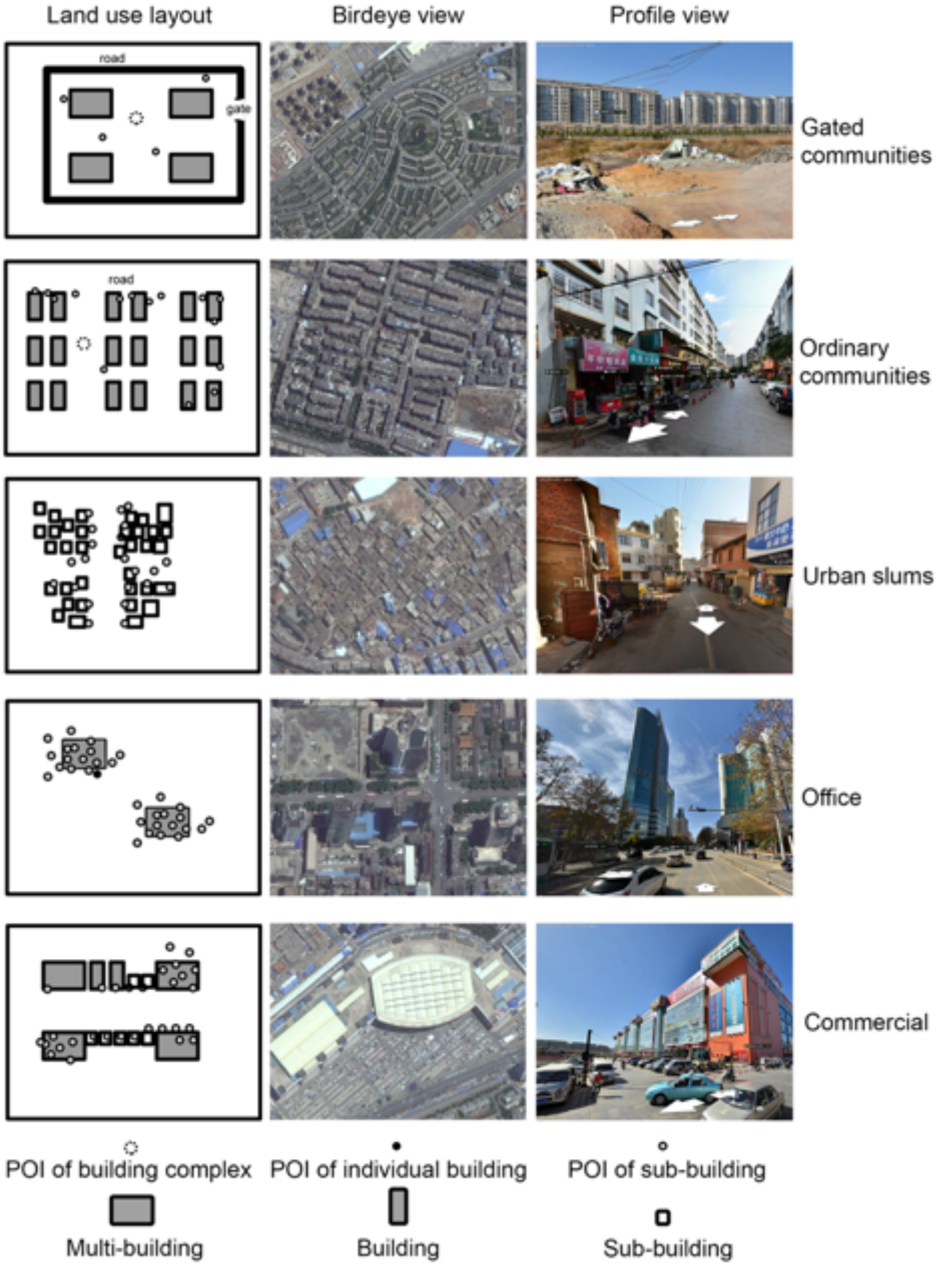


Figure 5. Primary Land Use in Kunming

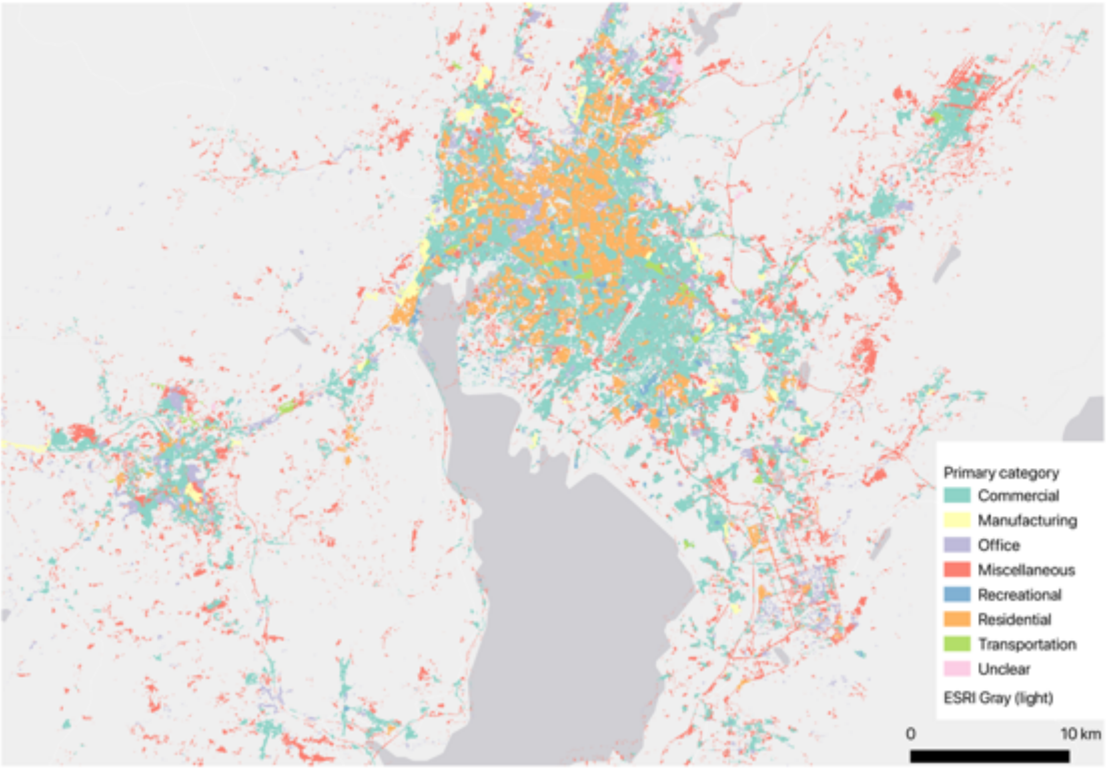
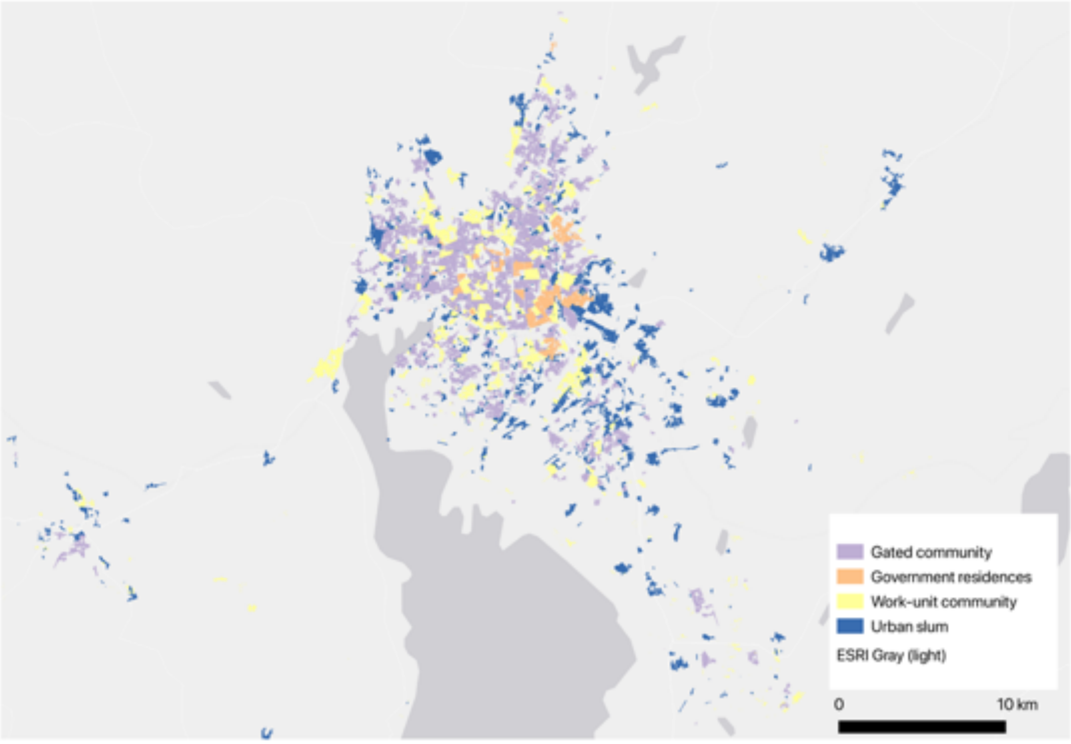


Figure 6. Urban Communities of Different Socio-political Statuses



SUPPLEMENTARY INFORMATION

Appendix A. Data Collection

I collect and store geotagged posts in real-time from citizens in Kunming on Sina Weibo, for a 10-month period from 1 November 2013 through 1 August 2014. Geotagged posts are those sent on a device, typically a smartphone, with a tag of location that a Weibo user chooses to show with the post, as well as to the underlying geographical coordinates from the GPS device on the smartphone at the time of posting. The location is highly precise: the positional error is 2 to 25 meters (Zandbergen and Barbeau 2011). The Sina Weibo Place Nearby Application Programming Interface (API) permits me, using a programming interface with location and search radius, to access the Sina Weibo database that stores all such posts.[25] To systematically retrieve posts in Kunming, I designed a net of approximately 102,165 location search points set 0.004 degrees or approximately 400 meters apart from one another; this entirely encompasses the boundaries of Kunming municipality.[26] The Sina Weibo Nearby API limits retrieving to a maximum of the 20,000 most recent posts from a single location point at any one time,.[27] I navigate my net of points, point by point, to retrieve the most recent posts within a radius I set at 1,000 meters with 23 API tokens and 6 computers.[28] By design, the radius of 1,000 meters for each location points set 400 meters apart from one another implies considerable spatial overlap, as illustrated below in Figure A.1. This contributes to intended redundancy in our retrieval of posts: in each two-week period, as I navigate our net of points systematically, from point to nearby point, at each point retrieving up to 20,000 most recent posts within the 1,000-meter radius, my retrieval covers every piece of Kunming's geographic space many times over.[29] Due to the large number of search points, it takes me approximately 10 days to two weeks to navigate every point in our net, retrieving posts at each point. However, the design allows me to miss a post only if there are more than 20,000 posts from some single location in two weeks—which is effectively impossible.[30] The data collection process results in 1,247,106 unique posts during the study period, as illustrated below

[25] The Sina Weibo Nearby API was deprecated in May 2017 when Sina tightened its API control.

[26] It encompasses the area from 24.29 to 25.27 degrees north in latitude and from 102.07 to 103.74 degrees east in longitude.

[27] Specifically, Sina Weibo Nearby API has a technical limit of 50 posts per page and limits retrieval to the most recent 400 pages. If we view all 400 pages from a single location point at the same time, which we do here, we can view a maximum of the most recent 20,000 posts.

[28] I have shared the Python script on a Github repository corresponding to this article. It can be found at https://github.com/placeasmedia/kunming_urban_communities.git.

[29] I subsequently eliminate multiple observations of the same post to create a dataset of unique posts.

[30] Empirically, I find this limit of 20,000 posts in the retrieval space of a single location point is reached in about three months.

in Figure A.2. I am confident that I have collected all geotagged posts in Kunming during the study period.

I have also collected POIs from Baidu Surrounding API (*zhoubian jiansuo*), Google Place Search API, and Tencent WebService Place API between May 2013 and July 2014. I then navigated repetitively the fishnet of search locations that are used to retrieve Weibo posts to search nearby POIs in the same study period. I compare the completeness, locational precision, and the categorical information of their POIs as discussed in the main text. In conclusion, I found that Tencent Map (previously known as Soso Map) has the most complete information, but also provides other kinds of complete spatial data such as road networks and street view images, which can be useful in this study. I then use POIs from Tencent and Baidu for this study. I subsequently collect POIs from Gaode through its POI Search API in 2015 and 2016.[31] Gaode has a considerably larger number of POIs in shops and services. But it essentially provides the similar amount of information on real estate communities and offices, and less information on the categories that are deemed sensitive, such as churches and mosques, than Baidu.

In addition, I have collected the road networks from Tencent through its Route Planning API (*luxian guihua*) in late 2013. To do so, I randomly place two points anywhere in Kunming and searched the routes between them, repeating this process until no new route can be found after the 100th iteration. I am confident that all the road network has been retrieved. Together with the road network that I download from OpenStreetMap, they form a complete set of road blocks in Kunming that can delineate almost all neighbourhoods (Figure A.3).

Furthermore, I collected the street view images from Tencent through its Street View Static Map API (*jiejing jingtai tu*). To do so, I use the collected POIs from Tencent as the search points. At each point, I collect four images that face 0, 90, 180, 270 degrees respectively at the pitch of -15 degree. According to Tencent, these street view images are acquired in 2013 and 2014 (https://zh.wikipedia.org/wiki/%E9%A8%B0%E8%A8%8A%E5%9C%B0%E5%9C%96#2013%E5%B9%B4), roughly corresponding to my study period. The collection results with 213,964 POIs with 539,527 street view images, which covers the most parts of Kunming.

Lastly, I collected the apartments in residential communities for sale from the real estate website Fang.com (formerly soufan.com), from July 2013 to July 2014 through web scraping. The listed properties range from the residential units in government compounds, such as the Yunnan Provincial Party School's dormitory, to nearly complete real estate projects, such as Dianchi Lingxiu phase one real estate. Together, there are 1,809 residential communities that have detailed records on the building types, prices, and so forth. A summary statistics is included as Table A.1.

[31] I share the Python scripts on the Github repository mentioned earlier.

Figure A.1. A Net of Points in a Sea of Posts

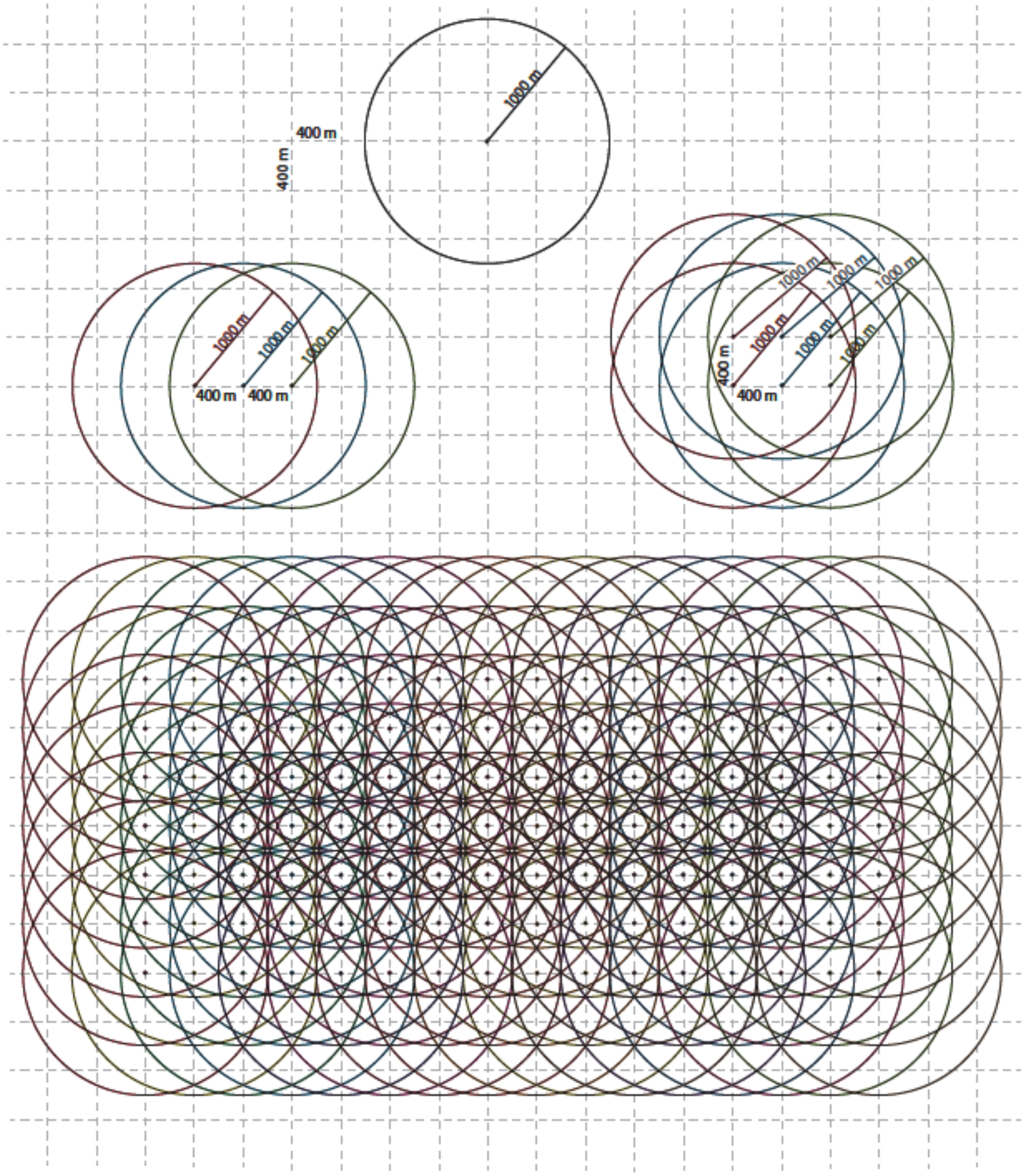


Figure A.2. Geotagged Posts in Kunming

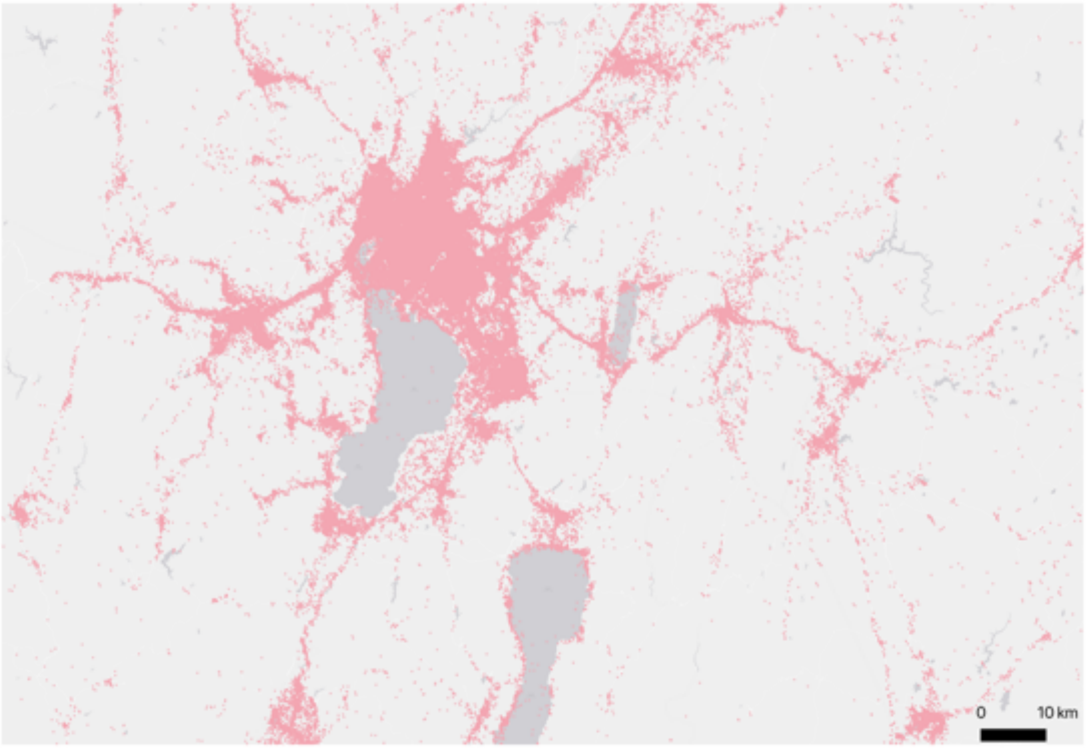


Figure A. 3. A Sample Area of Road Networks in Kunming

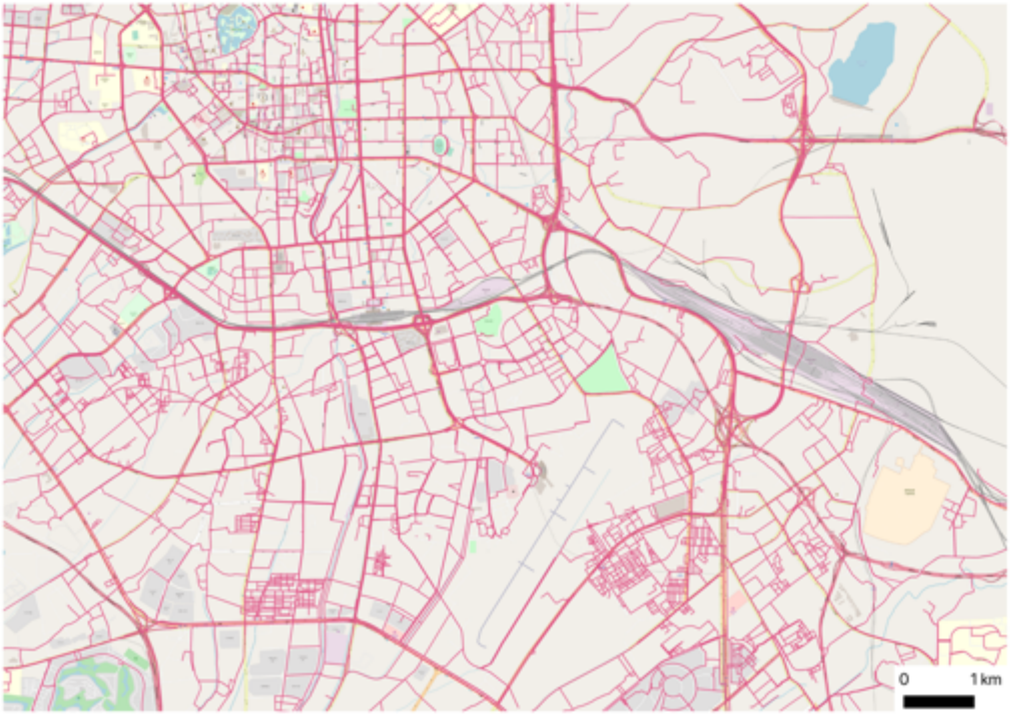


Table A.1 Summary Statistics of Housing Records in Kunming

Statistic	N	Mean	St. Dev.	Min	Max
Sale price (in Chinese Yuan)	1,393	18,404,758	365,343,558	150,000	8,710,000,000[1]
Rental price (in Chinese Yuan)	1,393	155,680	3,437,206	1,415	89,968,976[1]
Number of bedrooms	1,393	2.949	1.227	0	10
Number of living rooms	1,393	1.892	0.600	0	5
Unit area (in Square Meters)	1,393	127.895	73.563	17	730
Stories of the building	1,393	13.696	9.887	1	45

1. Some buildings are sold and rent as a whole, resulting a large amount of listing price.

Appendix B Sample Selection, Algorithms, and Evaluation in Land Cover Classification

I use all Landsat images during the study period as long as the estimated cloud coverage of the image is less than 50%. This results in 71 images in total, spanning from 1988 to early 2014 (Table B.1). To the extent that satellite images are less available in early years of the study period, I combine them at a multiple-year's interval in this period, from 1988 to 1992, from 1992 to 1995, and from 1996 to 2000. After 2000, the classification is performed at an annual basis.

Before mapping the land cover types, I need a training sample, a representative sample of pixels of each land cover type. Training samples have information of both the features and the labels of outcome. Features are the digital values in each spectrum associated with pixels in the training sample; labels of outcome are the correct land cover type for these pixels. The machine uses these samples to "learns" certain rules that can connect features with the correct type, or to put it with greater rigor, assign proper weights to different values of features so that they add up to a value that indicates the correct land cover type. The algorithm then applies rules so "learned" to pixels that are not in the training samples. If a pixel (900 m^2) contains mixed land cover types, the classified type is defined provisionally as whatever is made of over 50 percent of a pixel.

Manual selection of training samples is a time-consuming process since many pixels representing all targeted land cover types are called for; moreover, the training sample must be representative, accurate, and considerable in number so that sufficient information exists for machine learning algorithm to do its job. If the training sample lacks pixels that can distinguish the spectral difference between exposed soil and built-up areas, the machine will misjudge these two types. A high-quality training sample takes time to produce, since it requires researchers to manually select it by careful digitization.

To expedite the process, I use a semi-automatic method for sample selection. I use stratification to limit the number of candidate pixels from the entire image and then compute a Normalized Difference Vegetation Index (NDVI) in the summer of two periods, 1988-1993 and 2010-2013. NDVI provides strong indication that the land cover is vegetated.[32] I filter out areas that stay at high overall value in both periods as vegetated land and areas that stay at low value in both periods as non-vegetated land.[33] To select samples for urban land, I mask out non-vegetated land enclosed by small street blocks and select more than 213 individual pixels from it. Out of these small street blocks, the non-vegetated areas are my candidate sample pools (24 pixels) for exposed soil. The vegetated areas are my candidate sample pools for forest, which reveals itself in

[32] Donald W. Deering, DW, and Robert H. Haas, "Using Landsat Digital Data for Estimating Green Biomass", *NASA Technical Memorandum*, 80727 (1980), 1-21

[33] Gillian L Galford, John F Mustard, Jerry Melillo, Aline Gendrin, Carlos C Cerri, and Carlos EP Cerri, "Wavelet Analysis of MODIS Time Series to Detect Expansion and Intensification of Row-Crop Agriculture in Brazil", *Remote Sensing of Environment*, 112 (2008), 576–87

having relatively low seasonal variations of NDVI in each period. As for areas where NDVI variations are relatively high, they are my candidate for agricultural land. Using DEM, these distinctions can be further refined as follows. Slopes that range from 17 percent to 28 percent (15 degree to 25 degree) and lie at an altitude below the tree lines are forest, forest being mandated by Grain for Green policy.[34] Given this policy and its enforcement, I am justified in selecting a sample of 193 pixels from such slopes for stable forest. As for slopes gentler than those covered by forest, I select a sample of 43 pixels for stable grass. Areas that have extremely low values in NDVI (areas that are more or less flat) I select a sample of 70 pixels for water body.[35]

With all training samples selected, I am able to classify images into land patches of major cover types at a very fine spatial and temporal scale. Several supervised machine learning algorithms, including Support Vector Machine (SVM), Decision Tree (DT), Vanilla Neural Network (NN), and Maximum Likelihood (ML), are used in comparison. SVM, NN, and ML are programmed using ENVI and IDL, and DL is employed using C4.5 software. To compare the accuracy of these different supervised machine learning algorithms, I conduct a ten-fold cross validation using 80% samples as training set and 20% samples as testing set. The results of cross-validation show that on average SVM outperforms other algorithms (Figure B.1). To be sure, I select three sample areas to compare the accuracy of SVM against other popular supervised machine learning algorithms in Figure B.2. Three sample areas are selected to visually inspect the results across different machine learning algorithms. Panel A: the urban core Panel; B: a peri-urban area that experienced rapid urban expansion; Panel C: a new town. The results show that for ML errors do occur, especially when the satellite image is partly covered by cloud or when the scanner liner corrector-off (SLC) makes a slip. For DT and NN, the algorithms tend to over-estimate the land that has been developed into built-up areas.

Therefore, I use Support Vector Machine (SVM) to classify each image, SVM being more accurate than other supervised machine learning algorithms.[36] [37] To automatically eliminate misclassified pixels of built-up areas, I check the result of each image at anterior and posterior

[34] Ruth Sherman, Renee Mullen, Li Haomin, Fang Zhendong, and Wang Yi, "Spatial Patterns of Plant Diversity and Communities in Alpine Ecosystems of the Hengduan Mountains, Northwest Yunnan, China", *Journal of Plant Ecology*, 1 (2008), 117–36

[35] Chengquan Huang, Samuel N Goward, Jeffrey G Masek, Nancy Thomas, Zhiliang Zhu, and James E Vogelmann, "An Automated Approach for Reconstructing Recent Forest Disturbance History Using Dense Landsat Time Series Stacks", *Remote Sensing of Environment*, 114 (2010), 183–98

[36] Huang, C, LS Davis, and JRG Townshend, "An Assessment of Support Vector Machines for Land Cover Classification", *International Journal of Remote Sensing*, 23 (2002), 725–49

[37] Giorgos Mountrakis, Jungho Im, and Caesar Ogole, "Support Vector Machines in Remote Sensing: A Review", *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (2011), 247–59

time points. The next step is to integrate the results of images taken at different times of a year, a procedure that requires me to designate the land cover as built-up or non-built-up, depending on the most frequent land cover type among the results in that year. For instance, after integration, if a parcel of land is consistently non-built-up in January, March, and May, but converts to built-up in October, with all results correctly classified, it is designated non-built-up for that year. Lastly, to keep that larger parcel (>= 4500 m^2) intact, single pixels in adjacent years are annexed. Also, multi-temporal ensemble and majority voting are used for the post-classification adjustments so that built-up changes can be effectively captured, and noises and errors be removed.[38]

I take two steps to evaluate the accuracy of my land cover classification. My first step in evaluating the spatial and temporal accuracy of the mapped communities is to separate the urban core from the peri-urban area, the latter being the area that was converted from non-built-up to built-up after 1988 and can be monitored by remote sensing, supplemented by Google Earth Historical Images in high spatial resolution. To test for accuracy, I generate random points and compare them with available Google Earth Historical Images. I first generate 456 pixels randomly, stratified by the class of built-up area changes in my results (Figure B.3). I then use the Google Earth Historical Images and government planning maps to evaluate them.[39] From them I select points for evaluation at yearly basis or a few years interval when the original data are lacking, and I do so consistently across the built-up areas (Table B.2). In contrast to peri-urban, land cover change in the core area is from built-up to built-up, that is, one of urban renewal. Urban renewal change in the vertical direction cannot be tracked by Landsat images, and I am left to depend on dense road networks and street view images to evaluate its accuracy. By investigating the 53 randomly generated evaluation pixels that fall under the built-up areas in the downtown during the research period, my approach correctly classifies 48 pixels, including nine in which urban renewal occurred.

[38] Aguilar, Rosa, Raul Zurita-Milla, Emma Izquierdo-Verdiguier, and Rolf A De By. "A cloud-based multi-temporal ensemble classifier to map smallholder farming systems." *Remote sensing* 10 (2018): 729.

[39] The Google Earth Historical Images are available only from 1999 to 2014 in my study area. I rely on government planning maps to evaluate the changes in earlier period.

Table B.1. The Landsat TM/ETM Images Used in this Study

	1980s		1990s				2000s										2010s			
	88	89	92	93	94	96	00	01	02	03	04	05	06	07	08	09	10	11	12	13
Jan						6		21			6 30		3 19		9		30		20	6 22
Feb							20		9	28		9 25	20	23		12	15		5	7
Mar	30	9 25					23			16	2		8 24	11	29		3 19			11
Apr							24		30	9 25		6	1			17		15		20
May											21		19				6			22
Jun								14								4				
Jul																				
Aug			16																	
Sep							15					13		19						
Oct					25	30			7											
Nov							2	21		19			3 27		8	3 11			19	
Dec				24				23							10	29			21	

Table key:
23: Landsat 7 ETM scan line corrector-off

The images with large cloud coverage (>10%) were omitted. The top row indicates the year of the image that was taken, and the left column indicates the month of the image that was taken. The number in the table indicates the day of the image that was taken.

Table B.2. Number of Evaluation Points across Classes

Classes of land cover change	Number of evaluation points
Pre-1988	43
1988-1989	2
1989-1992	10
1992-1993	9
1993-1994	5
1994-1996	5
1996-2000	19
2000-2001	6
2001-2002	2
2002-2003	7
2003-2004	7
2004-2005	9
2005-2006	6
2006-2007	10
2007-2008	8
2008-2009	19
2009-2010	23
2010-2011	26
2011-2012	15
2012-2013	32
Agriculture	89
Forest	96
Waterbody	8

Figure B.1 The Comparison across Machine Learning Algorithms by Cross-Validation

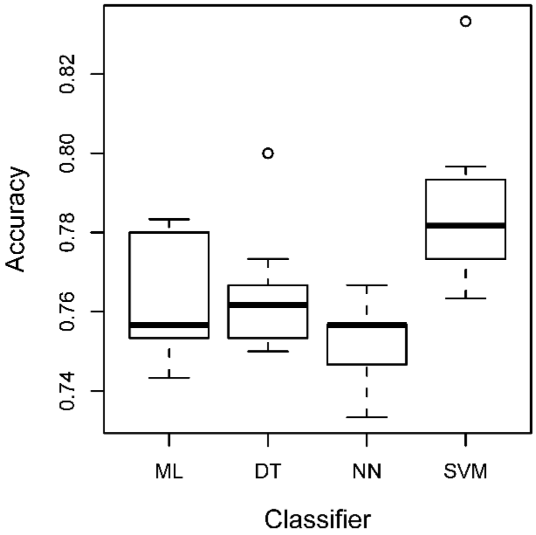


Figure B.2 A Visual Comparison of Three Exemplary Areas across Algorithms

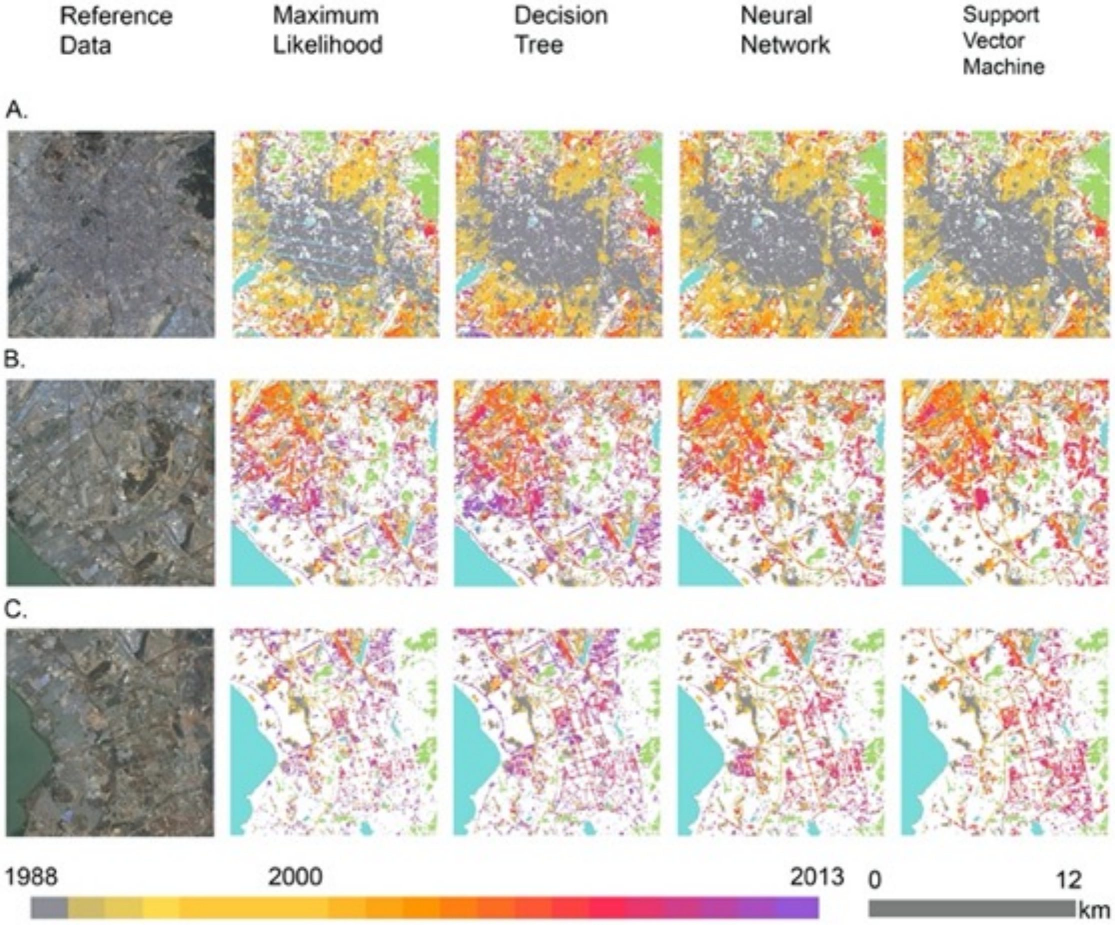
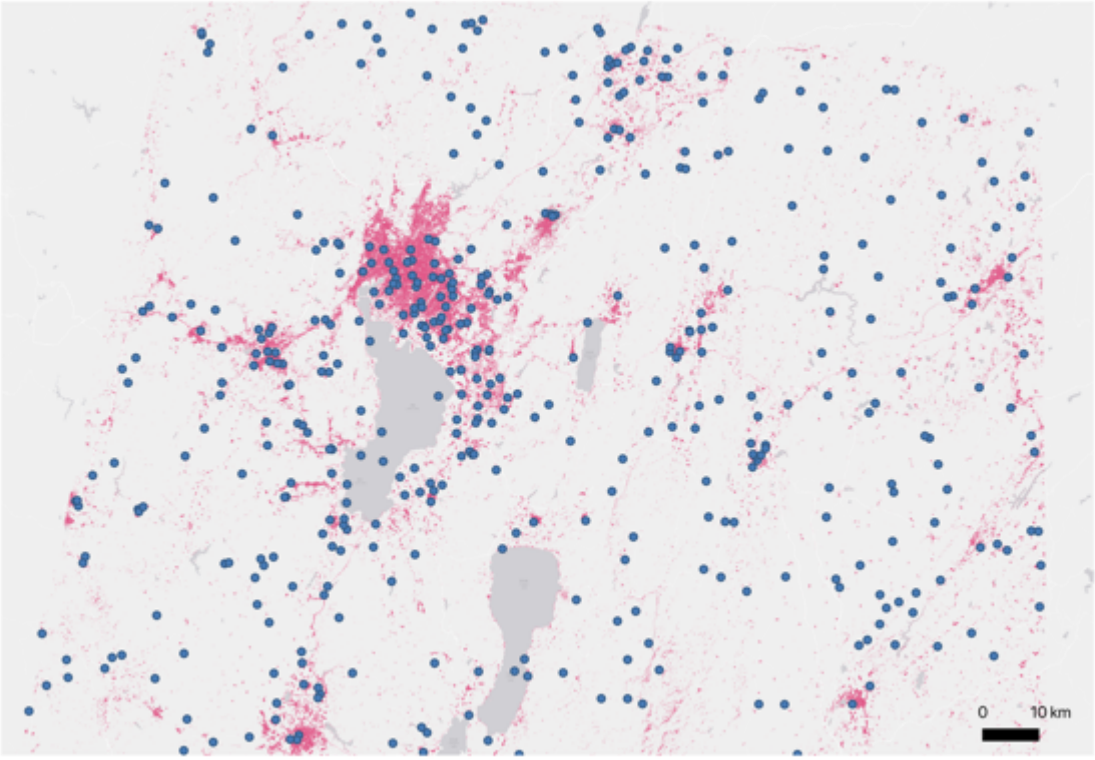


Figure B.3 The Spatial Distribution of Evaluation Points



Appendix C. Feature Generation and Machine Learning Algorithms in Land Use Classification

How many POIs are located on a parcel of land can be indicative to the primary use of the land. For example, an office building may house hundreds of companies at different sizes, while a manufacturing factory has very few POIs. To generate features to be used in the land use classification, I make use of the type and size of POIs. I code the types of POIs into the seven primary land use categories that I have defined in the Table 2, using the types of POIs defined by Tencent and Baidu in their metadata. This step includes the combinations of types that are described in the codebook of Tencent and Baidu mostly for their business interests. For example, a table provides a brief summary of the business types in POI defined by Baidu (http://lbsyun.baidu.com/index.php?title=lbscloud/poitags).

I also code the size of POIs in six categories. A massive infrastructure such as the airport is coded as 1, because one of POIs at this scale will naturally hosts hundreds of other POIs that indicate small businesses and shops. Next, large establishments such as manufacturing factories or parks are coded as 2. Then I code other smaller establishments accordingly, for example, universities as 3, museums as 4, libraries and gyms as 5, and shops as 6. An example of the codebook is included in Table C.1. I then calculate the number of POIs in various sizes and types on top of each land parcel.

Furthermore, I use the distance of a land parcel to POIs as training and testing features. Sometimes, a long distance to an important POI is more indicative of the land use of a parcel than a shorter distance to an ordinary POI. For example, a small land parcel that is more likely to be a part of an airport when it is 10 meters away from a food POI but 500 meters away from an airport POI. Thus, I code two binary variables to measure the distance of a land parcel to POIs of various sizes: whether a land parcel is within 150 meters of an office POI coded with size 3, and whether it is within 500 meters of a transportation POI coded as size 1.[40] In a similar way, I also code five binary variables to measure the distance of a land parcel to POIs of various land use types: first, whether a land parcel is within 150 meters of an office POI; second, whether it is within 250 meters of a recreational POI; third, whether it is within 500 meters of a transportation POI; fourth, whether it is within 150 meters of a manufacturing POI; and fifth, whether it is within 200 meters of a POI in miscellaneous category. The detailed list of features is included in Table C.1.

To select training and testing samples for the land use classification, I randomly distribute 559 points within the area that is covered in dense road networks or urban land from my first step. I identify them with the parcel of land and associate them with the features of land parcels, such

[40] Variables to measure the distance of a parcel to POIs of other sizes are highly correlated with other features and thus omitted.

as the number of POIs in primary categories and the POIs of various sizes on top of them. In the end, points with defined primary land use categories are used to evaluate my map: commerce, 67 points; office, 46 points; residential, 227; manufacturing, 53 points; transportation, 41 points; and recreation, 18 points. I then randomly split these points to a training sample that contains 80% of the points and a testing sample that contains 20% of the points, and then run a five-fold cross validation on them.

Before I classify the primary land use categories for all the land parcels, I compare popular supervised machine learning algorithms of their accuracy using the similar methods in my first step described in Appendix B. The machine learning algorithms tested are Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbours classifier (KNN), Decision Tree classifier (CART), Naïve Bayes (NB), Support Vector Machine (SVM), Gaussian Process Classifier (GPC), Random Forest (RMF). The results show that LR outperforms other algorithms in this classification task (Figure C.1). The classification report and confusion matrix are included in Table C.2, C.3, and Figure C.2. I then apply Logistic Regression algorithm to classify all the land parcels.

Table C.1. The Features in Land Use Classification

	Description
Year of development	The year of which the parcel of land is converted to the built-up areas.
Size of land	The spatial size of the land parcel.
Land use type defined by Tencent Map	Tencent map has a three-layer definition of the land use types. The first layer provides a broad category of land use types that depicts the primary land use in relative accuracy.
Distance to POI of size 1	The shortest distance to a POI of size 1.
Distance to POI of size 3	The shortest distance to a POI of size 3.
Distance to POI of recreational use	The shortest distance to a POI of recreational use.
Distance to POI of transportation use	The shortest distance to a POI of transportation use.
Distance to POI of manufacturing use	The shortest distance to a POI of manufacturing use.
Distance to POI of office use	The shortest distance to a POI of office use.
Distance to POI of miscellaneous use	The shortest distance to a POI of miscellaneous use.
Number of POIs at size 4 on top of the land	The number of POIs on top of the land parcel that has a coded size of 4.
Number of POIs at size 5 on top of the land	The number of POIs on top of the land parcel that has a coded size of 5.
Number of POIs at size 6 on top of the land	The number of POIs on top of the land parcel that has a coded size of 6.
Number of POIs of transportation use	The number of POIs on top of the land parcel that has a coded type of transportation.
Number of POIs of manufacturing use	The number of POIs on top of the land parcel that has a coded type of manufacturing.
Number of POIs of recreational use	The number of POIs on top of the land parcel that has a coded type of recreation.
Number of POIs of residential use	The number of POIs on top of the land parcel that has a coded type of residential use.
Number of POIs of governmental use	The number of POIs on top of the land parcel that has a coded type of governmental use.
Number of POIs of office use	The number of POIs on top of the land parcel that has a coded type of office use.
Number of POIs of miscellaneous use	The number of POIs on top of the land parcel that has a coded type of miscellaneous use.
Number of POIs of commercial use	The number of POIs on top of the land parcel that has a coded type of commercial use.

Table C.2. The Classification Report

	precision	recall	f1-score	support
Commercial	0.66	0.88	0.75	26
Manufacturing	0.33	0.33	0.33	3
Office	0.75	0.60	0.67	5
Miscellaneous	0.67	1.00	0.80	2
Recreational	1.00	0.67	0.80	6
Residential	0.92	0.80	0.86	41
Transportation	0.00	0.00	0.00	2
Accuracy			0.78	85
Kappa Score			0.66	85
weighted avg	0.79	0.78	0.77	85

Table C.3. The Confusion Matrix

	Commercial	Manufacturing	Office	Others	Recreational	Residential	Transportation
Commercial	23	0	1	0	0	2	0
Manufacturing	1	1	0	0	0	1	0
Office	1	1	3	0	0	0	0
Miscellaneous	0	0	0	2	0	0	0
Recreational	2	0	0	0	4	0	0
Residential	6	1	0	1	0	33	0
Transportation	2	0	0	0	0	0	0

Figure C.1. The Spatial Distribution of Training and Testing Points

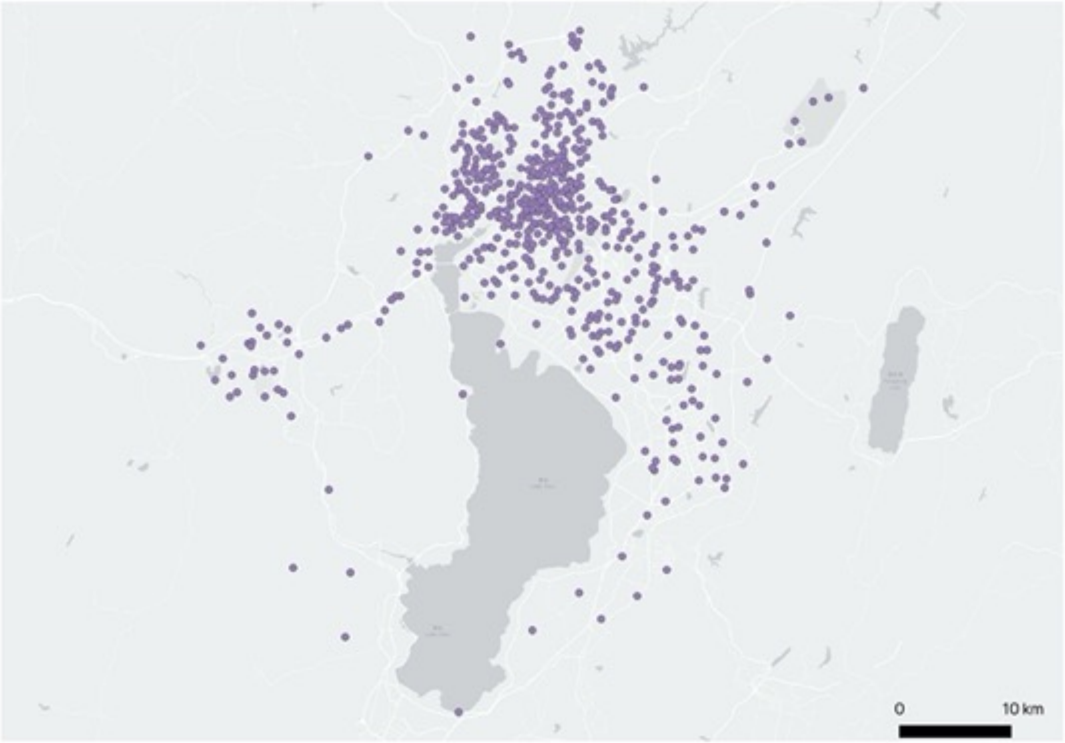
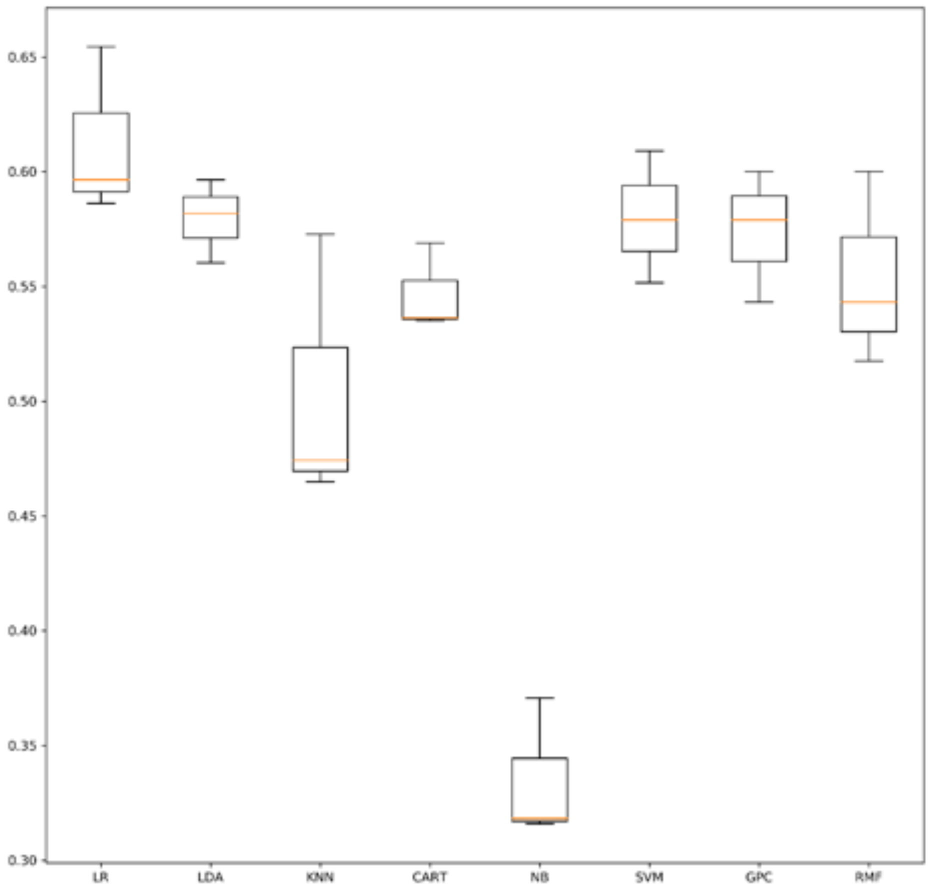


Figure C.2. The Comparison of Supervised Machine Learning Algorithms



Appendix D. Using Geotagged Social Media to Identify Urban Slums

To examine the temporal pattern of social media posts, I aggregate the geotagged posts in each urban neighbourhood during work hours and off-work hours. The work hours are defined as 9 a.m.- 12 p.m. and the off-work hour is defined as 11 p.m. – 2 a.m. Because many Chinese use smartphone and social media extensively before bedtime, these hours are particularly defined to exaggerate the differences between workplace and residences (Figure D.1). Moreover, the temporal pattern of social media posts is investigated on the Eve of Chinese New Year between 6 p.m. and 6 a.m., because most Chinese spend this time for family gathering (Figure D.2).

Sina Weibo data also come with rich information on the mobile devices that the Weibo user choose to show.[41] Is the price of mobile phone used on social media helpful in identifying different urban communities? To answer this question, I use a survey that was conducted in 2014 and 2015, only months after my study period.[42] Of 2,610 respondents, 1,296 reported using smartphone to access the internet. They were asked about their phone prices, which are coded in ordinally categorical variables. 1 stands for a phone below 500 Renminbi, 2 between 501 and 1,000 Renminbi, 3 between 1,001 and 1,500 Renminbi, 4 between 1,501 and 2,500 Renminbi, 5 between 2,501 and 3,500 Renminbi, 6 between 3,501 and 4,500 Renminbi, 7 above 4,500 Renminbi. They are also asked to provide an assessment of their living environment. 1 is much better than the average house in the city, which is presumably the apartment building block. 2 is slightly better. 3 is similar. 4 is slightly worse. 5 is much worse.

I plot the results in Figure D.3. It shows that residents have variegated preference in the phone price by different living environment. Nevertheless, those who live in a better environment (x = 1) appear to have significantly better phones than those who live in a worse environment (x=5). Because both variables are monotonically ranked, I run a Spearman's rank correlation coefficient test to see if these variables are correlated. The result indicates a significant negative correlation with Spearman's rho equal to -0.13. It shows that residents in rich communities do have better phones than those in poor communities, though the difference seems very small.

I then turn to the different urban communities in Kunming and the price of mobile devices used by residents who may live in these communities. Specifically, I identify the residents with their communities by examining from which type of urban communities they have dispatched a post during off-work hours.[43] I run regression with user fixed effects on the type of residences and

[41] There are more than 1,000 specific brands and models shown on Sina Weibo.

[42] The survey is a part of longitudinal survey, namely the Beijing Area Survey, which is conducted by Peking University. The survey has adopted rigorous methods and the responses are relatively reliable (see Shen, M., Yang, M., & Manion, M. (2010). Measuring Change and Stability over a Decade in the Beijing Area Study. In *A. Carlson, M. Gallagher, K. Lieberthal, & M. Manion (Eds.), Contemporary Chinese Politics: New Sources, Methods, and Field Strategies* (pp. 236-245). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511762512.017). A survey in Kunming is not practical due to the sudden attack at the railway station in 2014.

[43] The definition of off-work hours is 11pm – 2am on weekdays or anytime on weekend.

the price of their mobile device, using 161,525 identifiable observations. The results show that the price of their mobile devices is not significantly different across residential types (Table D.1). There could be many reasons for such insignificant differences, for example, residents who use cheaper phones may be less willing to post on social media, which may expose their phone models. I have also examined the average price of mobile devices on a map (Figure D.4), which confirms the use of mobile devices is highly mixed over urban neighbourhoods. Therefore, I exclude the price of mobile device in current analysis. The Confusion Matrix of the third step is included in Table D.2.

Table D.1. The Price of Residents’ Mobile Device and their Community Type

	Dependent variable:		
	Price of Mobile Device in RMB		
	(1)	(2)	(3)
Gated community (dummy)	1.490 (4.609)		
Work-unit community (dummy)		-3.578 (5.689)	
Urban slums (dummy)			8.201 (6.138)
Number of Observations	161,525	161,525	161,525
R2	0.891	0.891	0.891
Adjusted R2	0.810	0.810	0.810
Residual Std. Error (df = 92,695)	478.688	478.688	478.684
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table D.2. Confusion Matrix

	Gated community	Work-unit Community	Government Residence	Urban Slum
Gated community	47	6	3	0
Work-unit Community	32	24	0	1
Government Residence	0	0	10	0
Urban Slum	9	4	3	46

Figure D.1. The Average Differences in the Number of Social Media Posts between Morning and Night

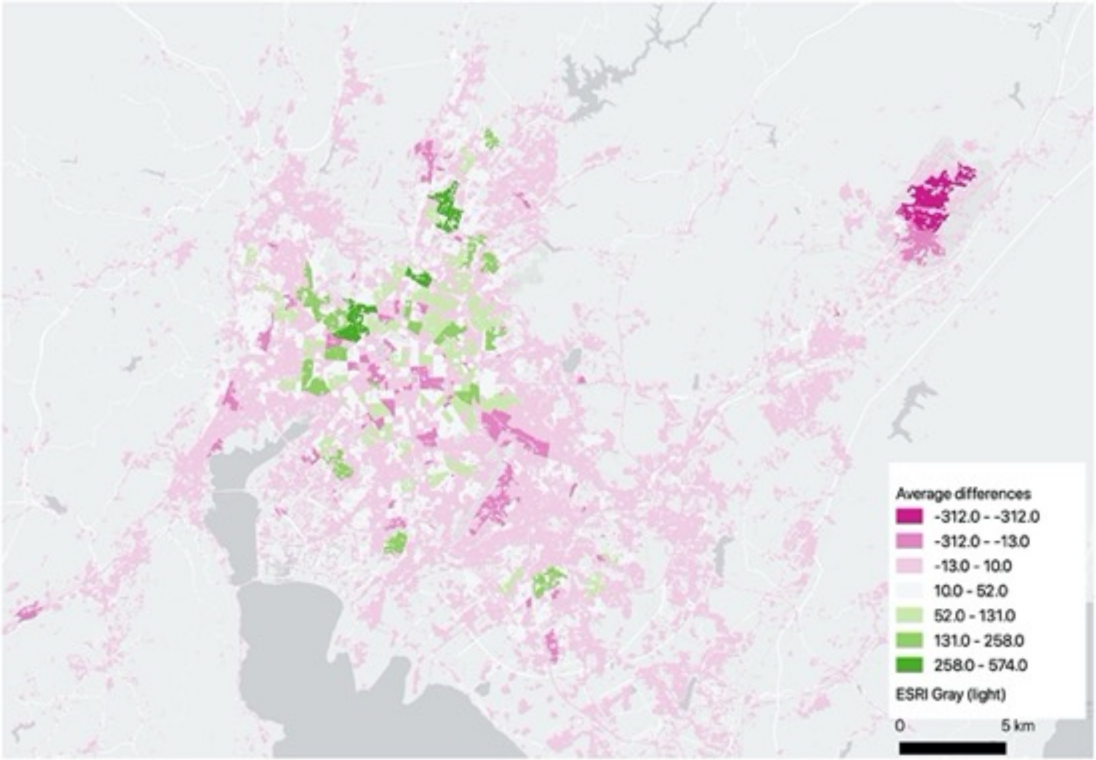


Figure D.2. The Number of Social Media Posts on The Eve of the Chinese New Year

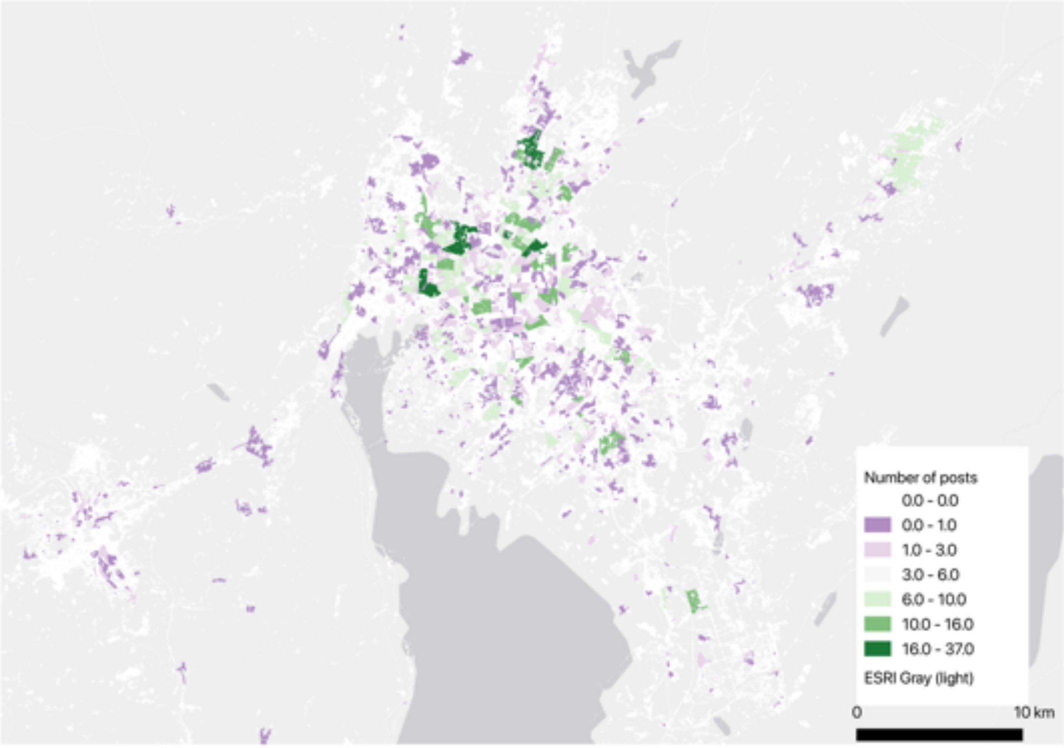


Figure D.3. The Plot between Mobile Device Price and Self-Reported Living Environment

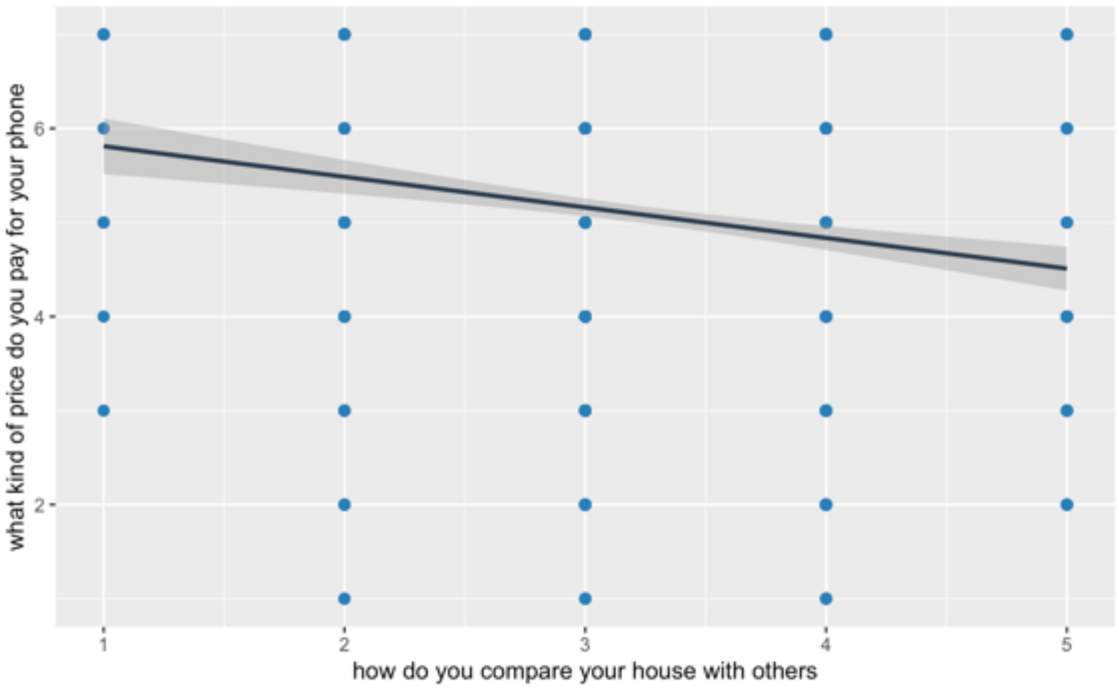
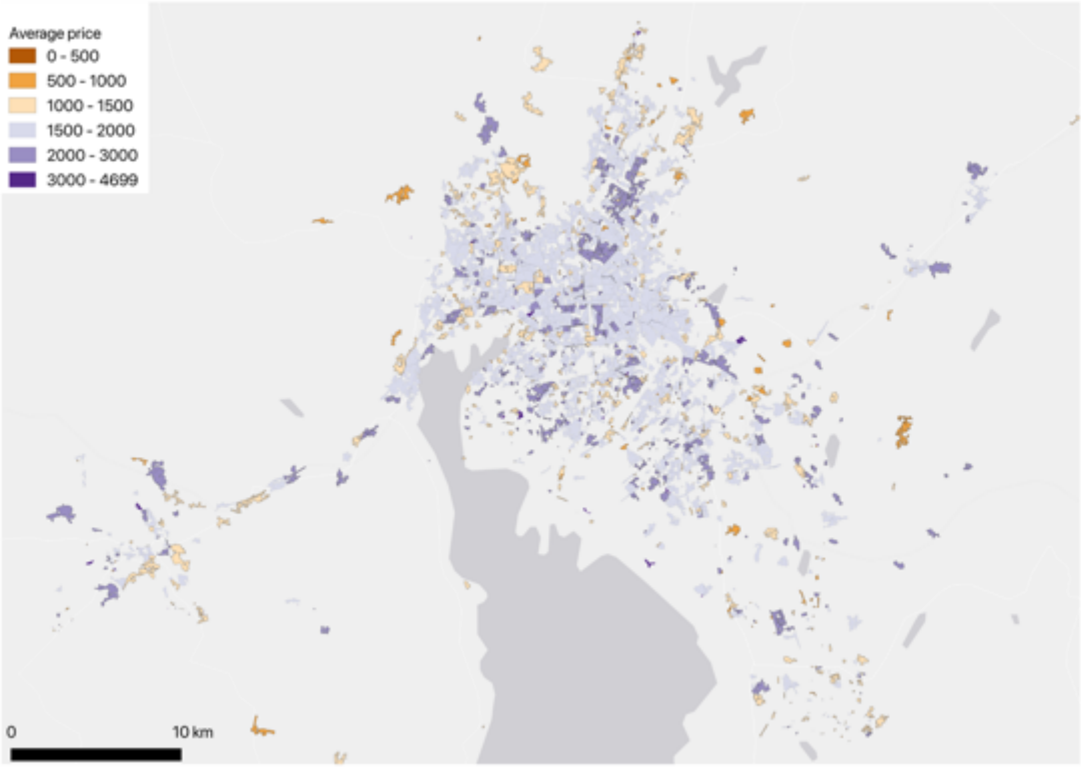


Figure D.4. The Average Price of Mobile Devices for Social Media Users


Average price
0 - 500
500 - 1000
1000 - 1500
1500 - 2000
2000 - 3000
3000 - 4699
0
10 km