**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.[[1]](#footnote-1) 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.[[2]](#footnote-2) 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,.[[3]](#footnote-3) 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.[[4]](#footnote-4) 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.[[5]](#footnote-5) 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.[[6]](#footnote-6) The data collection process results in 1,247,106 unique posts during the study period, as illustrated below 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. [[7]](#footnote-7) 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.

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

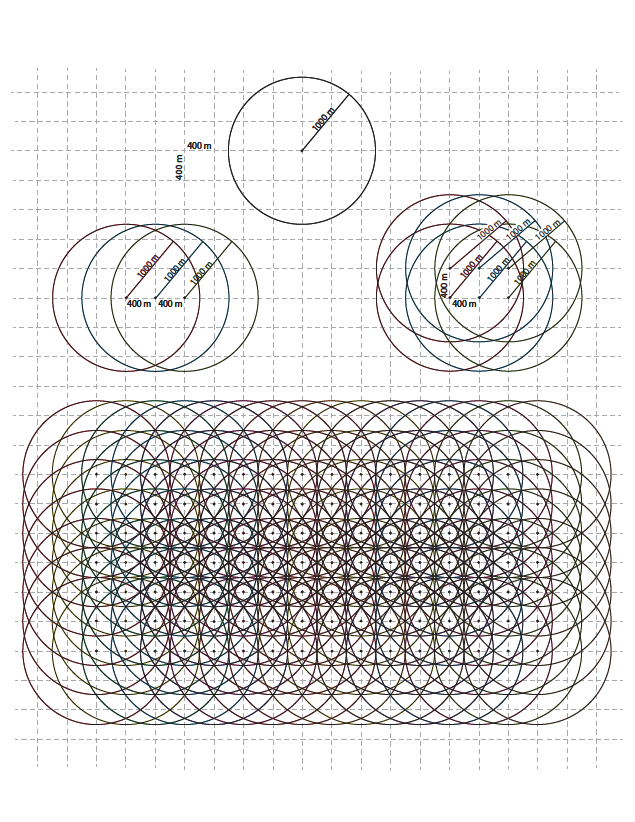


Figure A.2. Geotagged Posts in Kunming



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 | 8,710,000,0001 |
| Rental price (in Chinese Yuan) | 1,393 | 155,680 | 3,437,206 | 1,415 | 89,968,9761 |
| 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 m2) 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.[[8]](#footnote-8) 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.[[9]](#footnote-9) 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 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.[[10]](#footnote-10) 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.[[11]](#footnote-11)

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.[[12]](#footnote-12) [[13]](#footnote-13) To automatically eliminate misclassified pixels of built-up areas, I check the result of each image at anterior and posterior 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 m2) 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.[[14]](#footnote-14)

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.[[15]](#footnote-15) 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.

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

Fig_S3.tiff

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

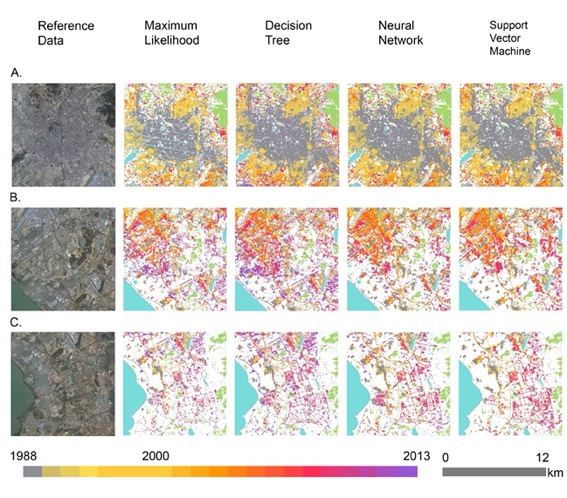


Figure B.3 The Spatial Distribution of Evaluation Points



**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.[[16]](#footnote-16) 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 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

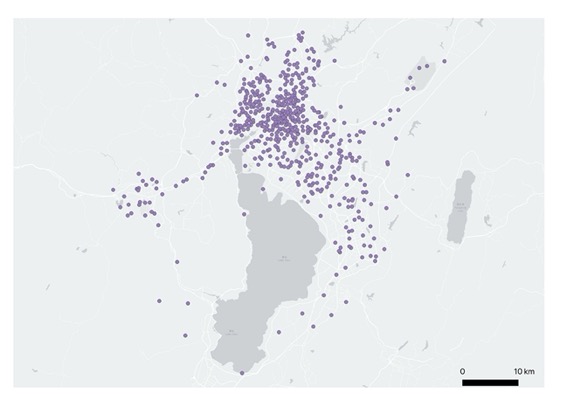
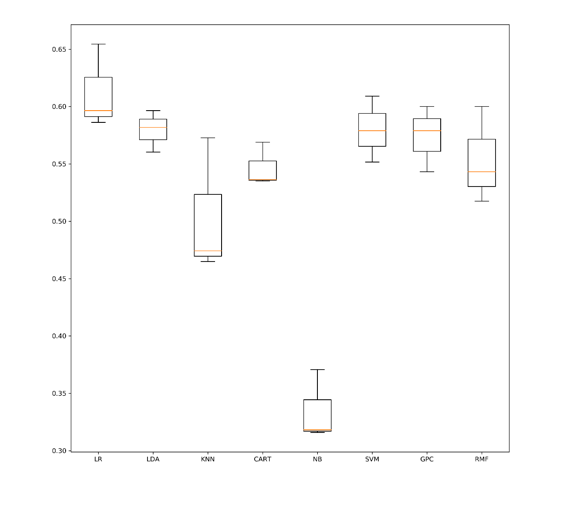


Figure C.2. The Comparison of Supervised Machine Learning Algorithms



**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.[[17]](#footnote-17) 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.[[18]](#footnote-18) 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.[[19]](#footnote-19) I run regression with user fixed effects on the type of residences and 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

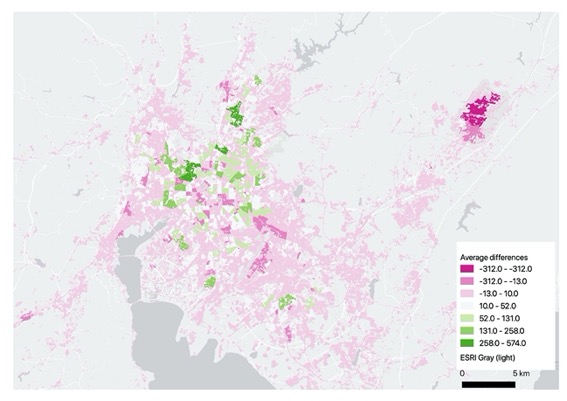


Figure D.2. The Number of Social Media Posts on The Eve of the Chinese New Year 

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

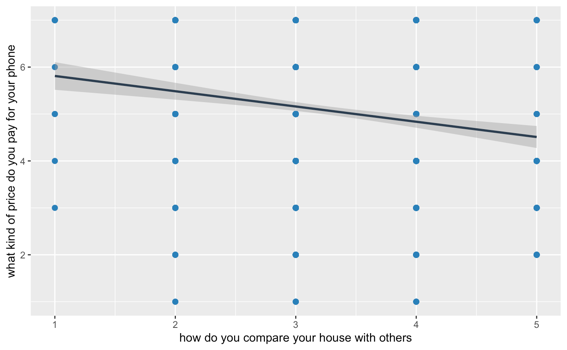


Figure D.4. The Average Price of Mobile Devices for Social Media Users



1. The Sina Weibo Nearby API was deprecated in May 2017 when Sina tightened its API control. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)
3. 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. [↑](#footnote-ref-3)
4. 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>. [↑](#footnote-ref-4)
5. I subsequently eliminate multiple observations of the same post to create a dataset of unique posts. [↑](#footnote-ref-5)
6. Empirically, I find this limit of 20,000 posts in the retrieval space of a single location point is reached in about three months. [↑](#footnote-ref-6)
7. I share the Python scripts on the Github repository mentioned earlier. [↑](#footnote-ref-7)
8. Donald W. Deering, DW, and Robert H. Haas, “Using Landsat Digital Data for Estimating Green Biomass”, *NASA Technical Memorandum*, 80727 (1980), 1-21 [↑](#footnote-ref-8)
9. 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 [↑](#footnote-ref-9)
10. 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 [↑](#footnote-ref-10)
11. 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 [↑](#footnote-ref-11)
12. 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 [↑](#footnote-ref-12)
13. 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 [↑](#footnote-ref-13)
14. 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. [↑](#footnote-ref-14)
15. 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. [↑](#footnote-ref-15)
16. Variables to measure the distance of a parcel to POIs of other sizes are highly correlated with other features and thus omitted. [↑](#footnote-ref-16)
17. There are more than 1,000 specific brands and models shown on Sina Weibo. [↑](#footnote-ref-17)
18. 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. [↑](#footnote-ref-18)
19. The definition of off-work hours is 11pm – 2am on weekdays or anytime on weekend. [↑](#footnote-ref-19)