Exploiting the Interdependency of Land Use and Mobility for Urban Planning

Kasthuri Jayarajah
Singapore Management
University
kasthurij.2014@phdis.smu.edu.sg

Andrew Tan
Singapore Management
University
andrewtan@smu.edu.sg

Archan Misra
Singapore Management
University
archanm@smu.edu.sg

Abstract

Urban planners and economists alike have strong interest in understanding the inter-dependency of land use and people flow. The two-pronged problem entails systematic modeling and understanding of how land use impacts crowd flow to an area and in turn, how the influx of people to an area (or lack thereof) can influence the viability of business entities in that area. With cities becoming increasingly *sensor-rich*, for example, digitized payments for public transportation and constant trajectory tracking of buses and taxis, understanding and modelling crowd flows at the city scale, as well as, at finer granularity such as at the neighborhood level, has now become possible. Integrating such understanding with heterogeneous data such as land use profiles, demographics, and social media, enables richer studies on land use and its interdependence on mobility. In this work, we share findings from our preliminary efforts and identify key lines of research inquiry that can help urban planners towards data-driven policy decisions.

Author Keywords

Urban computing, land use, urban mobility, clustering, prediction

ACM Classification Keywords

H.4 [Information Systems Applications]

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore.
Copyright © 2018 ACM ISBN 978-1-4503-5966-5/18/10$15.00.
https://doi.org/10.1145/3267305.3274163

Introduction

Urban planners are often posed with questions such as *what should the mix of use allocated to an area?* (e.g., how many restaurant lots), *what's the viability of a type of business if it's allocated the permit to operate in that area?*, *does the land use profile match the people flow to that area?* and so on. The series of Such questions can be reduced to a two-pronged problem entailing systematic modeling and understanding of how land use impacts crowd flow to an area and in turn, how the influx of people to an area (or lack thereof) can influence the viability of business entities in that area. Traditionally, planners have relied on surveys and limited observational studies to help them with such problems.

With cities becoming increasingly *sensor-rich*, there's opportunity now to comprehend the family of problems using large-scale, longitudinal data. For example, digitized payments for public transportation and the availability of GPS sensors aboard public buses and taxis enable the modelling of crowd flows at the city scale, and help further our understanding of commute patterns, disease spread, the impact of local events, etc. The fine spatio-temporal resolutions of data have made it possible to study variations across neighborhoods within the city. At the same time, the emergence and popularity of location-based social networks (LBSNs) such as Foursquare and Google Places have demonstrated their use in studying topics such as crowd flows and business longevity. In addition, they are also useful in profiling neighborhoods by their land use due to the rich semantics present in the user-generated data (e.g., type of cuisine a restaurant offers). Integrating such understanding with heterogeneous data such as land use profiles, demographics, and social media, enables richer studies on land use and its interdependence on mobility.

In this paper, we share our preliminary efforts and early findings in studying a specific question: *given the mix of land use in a neighborhood, can we estimate the utilization pattern of a "planned" carpark?*. To this end, we collected longitudinal car park lot availability data from over 1500 carparks from across Singapore, and extracted the land use around each individual carpark using social media, and predict the utilization pattern of a carpark using machine learning algorithms. Although we limit the scope to carpark utilization in the current work, the problem is generalizable to traffic and congestion levels in the neighborhood, in general, as it may be observable through taxi pickup/dropoff patterns, public transport utilization and so on.

We make the following key contributions in this work:

1. We first explore the use of unsupervised clustering in extracting *categories* of carparks based on their weekly, temporal utilization patterns. We share preliminary insights from the resulting clusters of carparks and the land use around those.
2. We then predict the utilization patterns of carparks, given the neighborhood land use mix, posing the problem as a multi-class classification task. We achieve an $AUC \approx 0.84$ suggesting that our primary hypothesis that the activity or land use of an area can be indicative of the people flow in that area.

Problem Description

In this work, we tackle the problem of predicting the temporal utilization pattern of a carpark given the land use around it. We adopt the following definitions and notations.

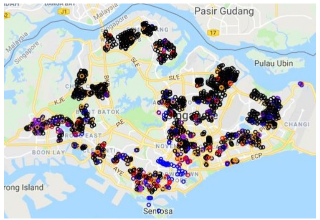


Figure 1: Carparks considered in this work. The 5 colors represent 5 different clusters resulting from the carparks' weekly temporal patterns. The dominant black carparks represent mostly carparks attached to residential blocks.

Given a carpark, c, at location $< latitude, longitude >$ representing its coordinates, we define its *weekly temporal profile*, $w_c(t) = mean(availability(c, t))$, where t represents a time window which is a combination of one of the seven days of the week and a observation bin representing the a slot during the day. For example, if the bin size is 1 hour, then the vector $w_c(t)$ is $7 \times 24 = 168$ in length. In this work, we consider bin sizes of 15 minutes. In order to reduce the problem space, and handle any noise in utilization data, we first cluster the weekly temporal profiles of all the carparks using $k-$ means, in an attempt to uncover *types* of carparks (we investigate the choice of k in the subsequent section).

Further, for each carpark, c, we define its neighborhood as the circular region encompassing a radius, r, from its $< latitude, longitude >$. We express this neighborhood, or, land use profile, as a distribution over venue categories, V (i.e., a vector of length $|V|$ whose elements are the counts of venues belonging to each category $v \in V$, within that radius r). Example venue categories are "Restaurants", "Apartments", "Office Spaces", etc.

Finally, we represent the multi-class classification problem with the dependent variable as the cluster type of carpark utilization and the input feature set, or the independent variables, being the distribution of venue categories around the carpark.

Datasets Used

In this paper, we share insights on the prediction of carpark utilization problem outlined in Section using the following categories of data, broadly. We summarize the datasets in Table 1.

Carpark Utilization: In order to understand people flow to an area, we rely primarily on information pertaining to public car park utilization; in particular, we use two sources for this purpose: (1) availability of lots at residential areas[1], i.e., at car parks attached to the Housing Development Board (HDB) which comprise of more than 70%[2] of the housing units across Singapore and (2) availability of lots in select commercial complexes in the greater Central Business District (CBD) of Singapore [3]. In total, we extracted longitudinal car park utilization data of 1699 car parks between April through July 2018.

As discussed earlier, we cluster carpark utilization patterns in an unsupervised manner, to uncover *categories* of patterns. With $k-$ means clustering (with k arbitrarily set to 5), Figure 1 shows the carparks belonging to each of the 5 clusters, differentiated by their color. As expected, as our dataset is skewed towards having a much higher number of carparks attached to residential blocks in the country, we see that the cluster represented in black is seen to be widespread and prevalent more dominantly.

As an illustrative example, we also plot the representative *centroid* weekly patterns of the five carpark clusters in Figure 2. The $x-$axis represents the time bins over a week (7 days × 96 fifteen-minute bins over each day). As

Dataset	Total Entities	Observation Period
Land Use	74,996 venues	as at July, 2018
Car Park Util.	1700+ carparks	April - July 2018

Table 1: Summary of datasets used in the analyses.

[1] https://data.gov.sg/dataset/carpark-availability

[2] https://www.straitstimes.com/politics/parliament-hdb-flats-made-up-73-of-singapores-total-housing-stock-in-2016

[3] https://www.mytransport.sg/content/mytransport/home/dataMall/dynamic-data.html

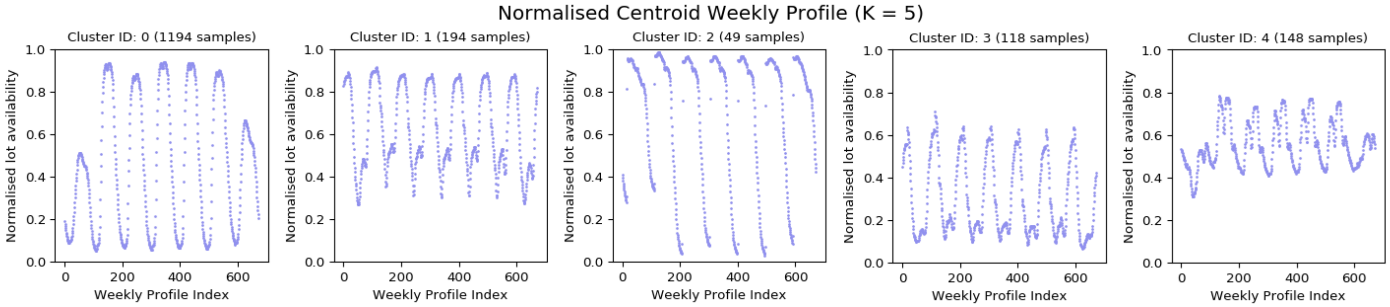


Figure 2: Weekly Carpark Utilization of the Cluster Centroids.

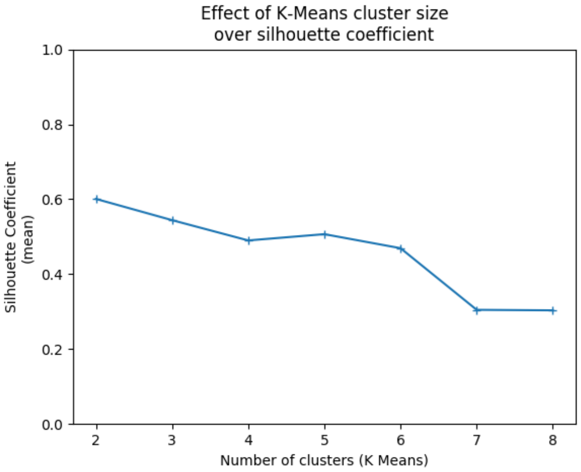


Figure 3: The impact of choice of the number of clusters k on clustering quality, measured by the Silhouette coefficient.

speculated, the cluster identified by "black" color in Figure 1 which is the same as the cluster identified as "Cluster 0" in this figure, aligns well with our expectations for residential carparks. The maximum lot availability is seen during the day hours (where the car owners would have left for their workplaces), and lowest availability is seen during the late evening hours (where the residents would have returned back home from work). Upon closer examination of the centroid clusters, it is also clear that "Cluster 1" and "Cluster 3" follow the expected patterns of commercial or office lots. Further investigations are warranted to understand the workings of the remaining clusters.

In order to choose the optimal k (i.e., number of clusters), we vary the number of clusters (from 2 to 8) and measure the Silhouette coefficient which is a measure of clustering quality. In Figure 3, we plot the number of clusters on the $x-$axis and the resulting coefficient on the $y-$axis. We conclude that for the current setting, $k = 2$ offers the highest quality and adopt this for the subsequent analyses.

Land Use: In order to extract the land use mix of neighborhoods (i.e., an area surrounding a given business entity), we rely on the Venue Search API [14]. Given a point location $< latitude, longitude >$, the API returns up to 50 venues within a configurable radius whose *gen* categories allow us to define the land use around that location. In this work, we set this distance to 500 meters as prior work [5] has shown that a venue's operation is affected primarily by conditions within this radial distance. We note here that alternative sources exist to extract such land use information, albeit at varying degrees of accuracy; for instance, the Google Places API [4] which functions similarly and returns up to 20 establishments within a neighborhood, whereas digitized land information from government authorities[5] are more likely to consist of the most accurate and fine-grained representation of a city's land use although harder to obtain. Figure 4 depicts the 75,000+ venues considered in this work.

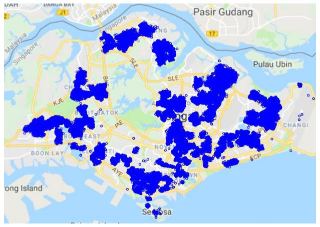


Figure 4: Foursquare venues considered in this work for profiling land use.

To further illustrate the validity of our central hypothesis that land use affects traffic patterns in the vicinity, we compare the utlization patterns (see Figure 5of two carparks, one from a residential area and the other from the Central Business District and the top-5 categories of venues around them, respectively (see Figure 6). In Figure 5, the patterns of the individual carparks clearly depict their similarity to the cluster centroids (residential: cluster 0 and CBD: cluster 1) with the times during which the lots fill or become free, being the inverse of the other. We observe that the top 5 categories of venues surrounding the residential carpark to be "Housing Development", "Apartment/Condos", "Parking", "Event Space", "Salon/Barber Shop". We note that it is common in Singapore for housing blocks to have their own allocated event space, or multifunction space, and neighborhood amenities such as barbershops. On the other hand, the top categories surrounding the carpark in the CBD area turned out to be, as expected, venues pertaining to the retail category: "Chinese Restaurant", "Sporting Goods Shop", "Clothing Store", "Miscellaneous Shops", "Japanese Restaurants".

Preliminary Results

In this section, we describe our early efforts in predicting the temporal utilization patterns of *planned* carparks.

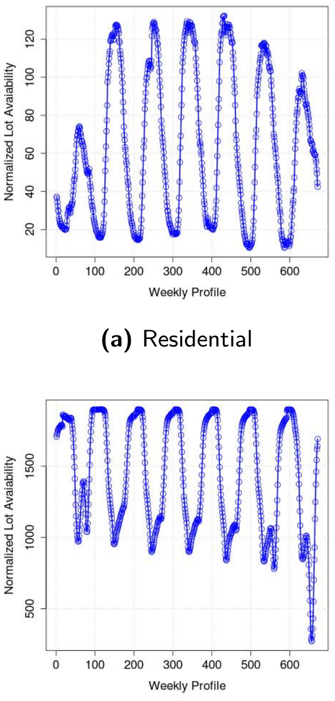
As described in earlier sections, we pose the problem of predicting the utilization patterns as a classification task where the class pertains to the unsupervised cluster or *category* of carpark. For each carpark instance, c, in the dataset, we consider the distribution of categories of Foursquare venues within a 500 m neighborhood as the input set of features. As a measure of feature selection, we prune out categories of venues that are not present around at least 100 of the carparks. This brings down the number of categories retained to 32 (from 506).

We pick the number of carpark clusters as $k = 2$ as it exhibits the highest Silhouette coefficient; in effect, this reduces the problem to a binary classification task. As our dataset consists of an unbalanced number of samples per class (i.e., Cluster 0: 1321 samples, Cluster 1: 378 samples), we first create a subset including all samples from cluster 2 and randomly sampled, equal sized samples from cluster 1, generating a balanced dataset. We then perform 10-fold cross-validation and report the average performance metrics in Table 2, for a number of machine learning algorithms.

We observe that the highest accuracy is obtained with

[4] https://developers.google.com/places/web-service/search

[5] https://www.sla.gov.sg/Services/Street-Map/Licensing-of-Digitised-Land-Information



(a) Residential

(b) CBD

Figure 5: The weekly utilization pattern two representative carparks, one from a residential area and the Central Business District area in Singapore.

Naive Bayes, with an AUC (area under the Receiver Operating Characteristic curve) of ≈ 0.84. The improved performance over other algorithms warrants further investigations which we defer as future work. We note that one possible reason could be either due to a lack of dependence between the features, or, the dependences either being distributed evenly across classes, or cancelling each other out.

Discussion and Ongoing Work

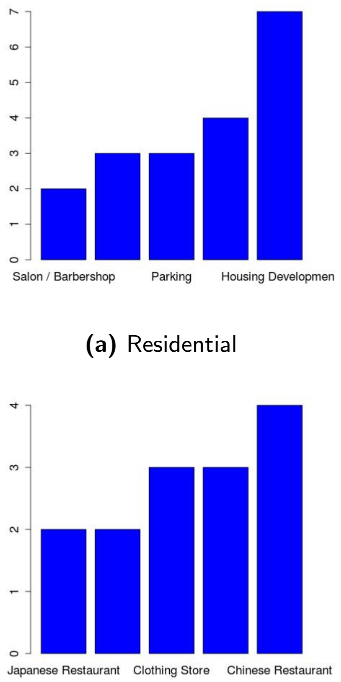
In this paper, we examined the fundamental hypothesis of whether and how the mix of land use in a neighborhood can impact the traffic level or people flow in that area. To operationalize this, we relied on social media for understanding the various business entities in a neighborhood (surrounding an existing or potential carpark), and weekly carpark utilizations as proxy for traffic or footfall to that area. In this section, we highlight some key limitations to the current work and outline lines of inquiry and open questions that we are actively pursuing.

Current Limitations: One of the key limitations in the current work lies with the data sources used in the analyses. For instance, the carpark utilization covers much of the residential carparks of the island, but only a few of the commercial carparks – as a result, the *categories* of carparks resulting from the clustering stage predominantly represent only two variations, and also suffer from skew in cluster sizes (i.e., one dominant, very large cluster, and several smaller clusters). Further, the business entities in the vicinity of a carpark are queried through the Foursquare API but are limited by the query results – i.e., there is a hard limit of only 50 results for every queries GPS location. For a fixed search radius, this means that the resulting search may not be able to distinguish between dense and sparse neighborhoods, which may lead to different results. In ongoing work, we are attempting to alleviate these drawbacks by (1) improving the coverage of existing datasets, (2) sourcing from alternate datasets (e.g., land use data from local planning agencies), and (3) consider additional layers of data (e.g., proxy for popularity of venues in the neighborhood using check-in data from social media, additional modes for learning footfall/traffic including taxi dropoffs, public transportation, etc.).

	Precision	Recall	F-Score	AUC
Naive Bayes	0.808	0.806	0.805	0.837
Random Forest	0.768	0.766	0.766	0.818
SVM	0.776	0.775	0.775	0.775
Logistic	0.704	0.704	0.704	0.752
Decision Tree	0.733	0.732	0.732	0.703

Table 2: Classification accuracy from using different machine learning algorithms.

Ongoing Work: In this current work, the key goal was to be able to predict the utilization patterns of a carpark based on the mix of entities surrounding it. As part of ongoing work, we are exploring the feasibility of being able to *explain* the utilization pattern of a transport resource (e.g., carparks, in this case) based on the activity or dynamics of the neighborhood. In other words, given a mix of activities (e.g., work, retail, residential, etc.), we would like to systematically estimate the expected traffic (e.g., road congestion) or utilization patterns as a generative process using technqiues such as tensor factorization. We believe that this would be a key enabler for urban planners in the decision making process of allocating new resources, or understanding the individual needs of neighborhoods. We are also actively investigating the impact



(a) Residential

(b) CBD

Figure 6: The top categories of venues in the neighborhood of two representative carparks, one from a residential area and the Central Business District area in Singapore.

of the spatial placement and function of other carparks in the vicinity – in other words, we believe that there are merits to considering the transportation resources as a multilayer network, instead of as isolated, independent entities.

Related Work

We briefly survey past literature in the following subdomains relevant to the work described in this paper.

Retail Performance and Land Use using Social Media: In recent years, the feasibility of using social media activity (check-ins in LBSNs, in particular) as proxy for understanding retail performance has been studied extensively. d'Silva et al. [7], the authors predict the expected demand patterns for a newly opening business entity based on its location and category of business by observing past patterns similar, existing businesses. Daggitt et al. [5] discuss the attractiveness of certain localities to categories of businesses. Further, Hristova et al. [11] investigate the spending behavior of consumers at sporting events whilst Karamshuk et al. [12] explore the use of social media for finding optimal placements for new retail stores. Additionally, works such as those of Zhao et al. [18], explain the economic impact of cultural investments (e.g., a new museum) on their surroundings.

Large Scale Urban Mobility: A large number of studies have explored the use of longitudinal, passively extracted data such as telecommunication records [2], taxi trips [16, 1], transit App logs [13], social media logs [10, 4], etc. for modeling urban mobility. These works have broadly focused on detecting work-home locations of the public [2, 4], commute patterns (e.g., for applications in optimizing trip time, efficient routing, etc.) [1], understanding disease propagation [17], forecasting traffic patterns [16] and detecting the occurrence of anomalous events [3, 15, 13], etc. In contrast, the nature of problems described in this paper attempt to understand the interplay between the two facets, mobility (extracted and learned from longitudinal traces similar to these works) and land use/activity.

Interplay between Mobility and Urban Activity: A few recent works have focused on establishing the interdependency between mobility and urban activity, at various levels. More close to our work, recently, d'Silva et al. [6] explore the impact of factors such as the catchment of a neighborhood, connectivity or accessibility of a neighborhood and the alignment of a business establishment with its neighborhood, on the longevity of businesses in an area. Further, Georgiev et al. [9] discuss the indirect impact of events (which draw unusually large number of people to an area) on the profitability of retail businesses in the area. Fan et al. [8] propose and evaluate a theoretical framework that factorizes mobility patterns of an area as a mixture of various activity patterns (e.g., work, commute, etc.) conducted in that area. The activity patterns, in effect, are artefacts of the nature of land use of that area. Similar in spirit to these works, we attempt here to understand the interplay between urban land use and mobility, in a broader sense.

Concluding Remarks

In this work, we posed the problem of interdependency of land use and mobility and shared our preliminary results from tackling the specific question of estimating the carpark utilization patterns based solely on the land use profiles around it. We show that even with a simplistic representation of the problem, a reasonably high accuracy of 0.84 (AUC) can be achieved validating our key hypothesis. We discussed limitations of the current work and outlined our next steps in studying this problem in-depth.

ACKNOWLEDGMENTS
This material is supported partially by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centers in Singapore Funding Initiative. K. Jayarajah's work was supported by an A*STAR Graduate Scholarship.

REFERENCES

1. Rajesh Krishna Balan, Khoa Xuan Nguyen, and Lingxiao Jiang. Real-time Trip Information Service for a Large Taxi Fleet *(MobiSys '11)*.
2. Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2013. Human mobility characterization from cellular network data. *Commun. ACM* 56, 1 (2013), 74–82.
3. Sanjay Chawla, Yu Zheng, and Jiafeng Hu. Inferring the Root Cause in Road Traffic Anomalies *(ICDM '12)*.
4. Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and Mobility: User Movement in Location-based Social Networks *(KDD '11)*.
5. Matthew L. Daggitt, Anastasios Noulas, Blake Shaw, and Cecilia Mascolo. 2016. Tracking urban activity growth globally with big location data. *Open Science* 3, 4 (2016).
6. Krittika D'Silva, Kasthuri Jayarajah, Anastasios Noulas, Cecilia Mascolo, and Archan Misra. 2018. The Role of Urban Mobility in Retail Business Survival. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (Sept. 2018).
7. Krittika D'Silva, Anastasios Noulas, Mirco Musolesi, Cecilia Mascolo, and Max Sklar. 2017. If I build it, will they come? Predicting new venue visitation patterns through mobility data. In *SIGSPATIAL*.
8. Zipei Fan, Xuan Song, and Ryosuke Shibasaki. 2014. CitySpectrum: A Non-negative Tensor Factorization Approach *(UbiComp '14)*.
9. Petko Georgiev, Anastasios Noulas, and Cecilia Mascolo. 2014. Where Businesses Thrive: Predicting the Impact of the Olympic Games on Local Retailers through Location-based Services Data *(ICWSM 2014)*.
10. Samiul Hasan, Xianyuan Zhan, and Satish V Ukkusuri. 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*.
11. Desislava Hristova, David Liben-Nowell, Anastasios Noulas, and Cecilia Mascolo. 2016. If You've Got the Money, I've Got the Time: Spatio-Temporal Footprints of Spending at Sports Events on Foursquare. (2016).
12. Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement *(KDD '13)*.
13. Tatsuya Konishi, Mikiya Maruyama, Kota Tsubouchi, and Masamichi Shimosaka. 2016. CityProphet: City-scale Irregularity Prediction Using Transit App Logs *(UbiComp '16)*.

14. n.d. 2018. Search for Venues. `https://developer.foursquare.com/docs/api/venues/search`. (2018). [Online; Last accessed 09-May-2018].

15. Linsey Xiaolin Pang, Sanjay Chawla, Wei Liu, and Yu Zheng. On Mining Anomalous Patterns in Road Traffic Streams *(ADMA'11)*.

16. Masamichi Shimosaka, Keisuke Maeda, Takeshi Tsukiji, and Kota Tsubouchi. 2015. Forecasting Urban Dynamics with Mobility Logs by Bilinear Poisson Regression *(UbiComp '15)*.

17. Yingzi Wang, Xiao Zhao, Anastasios Noulas, Cecilia Mascolo, Xing Xie, and Enhong Chen. Predicting the Spatio-Temporal Evolution of Chronic Diseases in Population with Human Mobility Data *(IJCAI'18)*.

18. Xiao Zhou, Desislava Hristova, Anastasios Noulas, Cecilia Mascolo, and Max Sklar. 2017. Cultural investment and urban socio-economic development: a geosocial network approach. *Open Science* 4, 9 (2017). `DOI: http://dx.doi.org/10.1098/rsos.170413`