The Workshops of the Tenth International AAAI Conference on Web and Social Media
CityLab: Technical Report WS-16-16

Capturing Real-Time Public Space Activity Using Publicly Available Digital Traces

Kostas Cheliotis
The Bartlett Centre for Advanced Spatial Analysis, University College London
Gower Street, London W1E 6BT, United Kingdom
kostas.cheliotis.12@ucl.ac.uk

Abstract

The study of public space activity has been one of the main foci in the debate on urban space transformations for the past decades, with many researchers adding to the debate through theoretical work, as well as empirical/quantitative evidence, drawing from their direct observations of public space activity. This paper attempts to enhance this approach of urban research and public space observation, by investigating the application of remote sensing techniques in public space analysis. More specifically, it attempts to capture public space activity using publicly available digital traces, such as environmental and temporal data, as well as social media data streams. By applying bivariate and multivariate analysis techniques to these datasets, it illustrates the possibility of capturing current activity in public spaces, with some degree of confidence. Furthermore, given the ubiquitous and real-time nature of these datasets, it also becomes possible to provide continuous estimates as well as short-term predictions on current and near-future public space use. Finally, it outlines the capabilities of this approach, to be used in complementary fashion to direct observation methods mentioned above, in building high resolution models and simulations of public space activity.

Introduction and Aims

Public Space is one of the main topics in the general discussion regarding issues of urban space, and has been for some decades now, often noted by scholars as having been introduced to the mainstream discussion by Jacobs' 1961 work 'The Death and Life of Great American Cities'. Half a century later, it is generally agreed upon that public spaces, along with their subsequent use, play a vital role in the overall image and cohesion of the city they represent, even though it is still debated what that role is or should be, as illustrated in recent reviews on the topic (Carmona 2010a; 2010b).

Nevertheless, researchers have for the past decades been studying the various facets of public space, oftentimes through direct observation and documentation. Approaches range in focus from pure morphology and typologies of public urban space (Krier 1979), to design function classifications (Gehl and Gemze 2000), to the *'way users engage with space'* (Dines, Catell, and Gesler 2006, in (Carmona 2010b)). In a more proactive approach, relating to the design of successful public spaces, researchers have surveyed and documented the usage of existing successful public spaces in order to understand what constitutes a successful space (Whyte 1980), by focusing on the individual activities taking place as well as the interactions between the users (Whyte 1988; Gehl 1987). It is these latter approaches that are of interest here, as they approach the matter from a user-centric perspective, identifying use and interaction as the main processes that make for a good public space, and can offer valuable insight in understanding the use of space as a bottom-up approach.

Given the usefulness of such approaches, it is unfortunate that this method of observation has some inherent limitations. As it is based on direct observation, it requires researchers to be passively present at the area of interest, or at the very least the installation of infrastructure such as cameras and recording devices, the latter carrying other issues in their use. Such requirements limit the gathered data to the very specific time ranges during which a space was documented. This results in very high quality data for surveyed periods, but unfortunately no data can be inferred for unsurveyed periods. This paper approaches this shortcoming in the context of contemporary big data, and offers potential solutions to this limitation.

Townsend (2000) argued that the increasing use of mobile phones signalled the arrival of the *'real-time city, in which* [urban] *system conditions can be monitored and reacted to instantaneously'*. Furthermore, with the digital traces generated today through the use of such new devices, it is becoming increasingly possible to capture the micro-interactions which make up the collective entity that is a city. Following in this vein, additional approaches have demonstrated the capabilities of urban data in monitoring and visualizing real-time urban activity as it is comprised from individual traces (Calabrese et al. 2011; Ratti and Claudel 2014), and further employing such data sets in inferring urban activity at the individual level (Diao et al. 2015). It is hypothesized that due to the portability and ubiquitous nature of mobile devices, within the next 20 years most urban data will be sourced from digital sensors, and will be available in various forms, with temporal tags as well as geotags in many instances (Batty et al. 2012).

Based on the advent of ubiquitous and real-time data availability, this paper identifies a potential opportunity in using Real-Time Data (RTD) to provide continuous, real-time estimations of current public space use. More specifically, the approach presented here acts as complementary to observational/empirical data, allowing for current activity to be captured without the need for in-situ recordings, and further infer current activity (and subsequent quality of public space) using remote sensing and data mining techniques. It is hypothesized then that by analyzing the effect environmental and other conditions have on the digital traces of public space users' activity, as captured by remote sensing and data mining techniques, we can begin to form a preliminary model for continuously capturing and subsequently predicting public space use.

Data and Methodology

A major municipal green space in London, UK (Hyde Park) is chosen as a location for a London Living Labs (L3) project[1], investigating novel uses of real-time data and networked infrastructure in managing and experiencing urban parks (ICRI 2015). Data presented in this paper is part of an on-going case study within the L3 context, which analyzes visitor activity in Hyde Park. It focuses exclusively on publicly available data, such as data released under a free or open licence, and data made publicly available by its authors (eg. users' public posts in social media platforms). An attempt is thus made to build a real-time profile of current activity in the area of interest, based on ambient geospatial information (Stefanidis, Crooks, and Radzikowski 2011) gathered from a variety of sources. This paper presents initial findings from the analysis of collected data, covering a period of 134 days, from the 14th of September 2015, up to (and including) the 27th of January 2016, with the aim of later using these findings in Real-Time applications.

Although the datasets discussed here comprise of archived data, this is simply for the sake of analysis. Dynamic (Real-Time) Data, including records retrieved from social media platforms and weather forecasts, contains datasets which provide information on current activity, and are retrieved and updated in real-time. More specifically, posts in social media platforms Twitter and Instagram are used as a proxy of current on-site activity. This data is collected regularly using the respective platforms' search APIs[2], via automated scripts written in the Python programming language, similar to (Hawelka et al. 2014). Weather forecasts are similarly retrieved using custom python scripts from forecast.io, in order to have an automated structure of current weather conditions. Planned events and gatherings in the area of interest in the near future are retrieved using Facebook's API, using the search term 'Hyde Park' and filtering events taking place in London, UK, and are used as an indicator of expected increased activity. Finally, in order to confirm and further validate correlations between remotely sensed data and in-situ activity, as well as for further calibration, park visitor counts were conducted via manual counts at the area of interest, at specific dates and times.

Data in most cases is formatted at two levels. At the aggregate level, data is presented as daily summary totals, depending on dataset. At a finer level, data is presented in more detail as an hourly total. Social media data is further cleaned up to remove duplicates (multiple posts from the same user within a certain period of time of 30 minutes), and only includes geotagged posts originating from within the park. Regarding weather data, different indexes were captured in the data collection, including temperature (minimum and maximum for daily summary), cloud coverage, precipitation probability and intensity, and wind speed. Planned events captured via Facebook's API further included start and end date and time, as well as the number of event attendees, as captured by the platform.

Regarding the methodology, these datasets are examined using standard bivariate and multivariate linear regression approaches. Social Media posts as captured from Twitter and Instagram are used as a proxy for visitor activity, and are considered the dependent variable in all cases, where correlation is investigated between Social Media Posts on the one hand, and climate/temporal attributes on the other.

Exploratory Data Analysis and Data Cleanup

Some initial characteristics and general properties of the dataset can be seen by looking at the raw data overview, with social media posts (SOCM) values shown in daily totals for the whole duration of 134 days (Figure 1). Values vary greatly, from the low hundreds to almost 5000, averaging 1068 daily SOCM, with a long-tailed distribution toward higher values. Zero values indicate collection failure, days where the automated collector scripts were not executed properly, and thus no data was captured for that day.

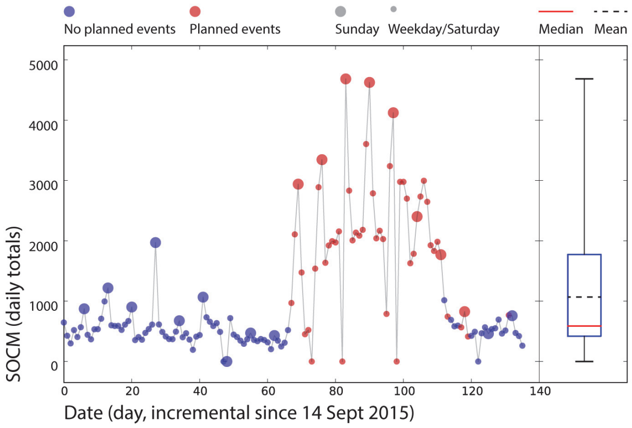

Figure 1: SOCM Time Series

Although the dataset presents a fairly scattered distribution, additional filters allow for a more meaningful interpre-

[1] L3 is a collaborative effort between research institutions and stakeholders, including UCL, Intel ICRI Cities, the Royal Parks, the Future Cities Catapult and Imperial College London, pursuing innovative ideas and applications within the Smart Cities context.

[2] https://dev.twitter.com/rest/public,
https://www.instagram.com/developer/endpoints/

tation of the dataset. Days containing planned events are highlighted (shown in red in Figure 1), overlapping with the majority of the highest recorded daily SOCM values. Indeed, most of these dates correspond to a major winter festival that takes place in the area of interest every year, attracting thousands of visitors, over a period of 45 days. Additionally, datetime information has been embedded in the dataset, and individual days of the week have been codified as integers (dayInt: Sunday to Saturday, 0-6 respectively). Sundays are then highlighted (larger point size, Figure 1), further corresponding to high SOCM values, both for event and non-event days. In annotating Sundays, a periodic characteristic of the dataset can be identified, where Sundays generally highlight the peak at each 7-day period, compared to fairly equal values throughout the rest of the week.

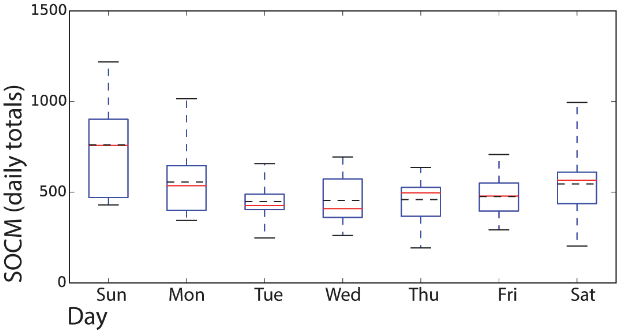

Figure 2: SOCM by Day of the Week

The aim of this analysis is to investigate the effect of environmental and temporal characteristics on public space use (measured as social media posts) during normal conditions. In this context, days with planned events are considered known outliers, with artificially high values. As such, days with planned events, along with zero value days (failed recordings), will not be considered for the rest of this analysis, as these records would introduce a strong bias. Even having removed known outliers, increased activity on Sundays is further evident when comparing SOCM by day (Figure 2). Most SOCM are recorded during Sundays, averaging 750 daily total, with values falling sharply on the next days, and picking up again on Saturdays.

Analysis

This section will be looking at the effect that climate and temporal characteristics have on recorded social media posts, first at the daily aggregate level, and later at an hourly level.

Daily Aggregate

At a daily aggregate level, initial assumptions focused on temperature being the main driver of park visitor activity (and thus social media activity), stating that days with higher temperatures would attract higher visitor numbers. This turned out to be a false hypothesis, as can be seen on Figure 3, showing daily SOCM levels against daily minimum and maximum recorded temperatures.

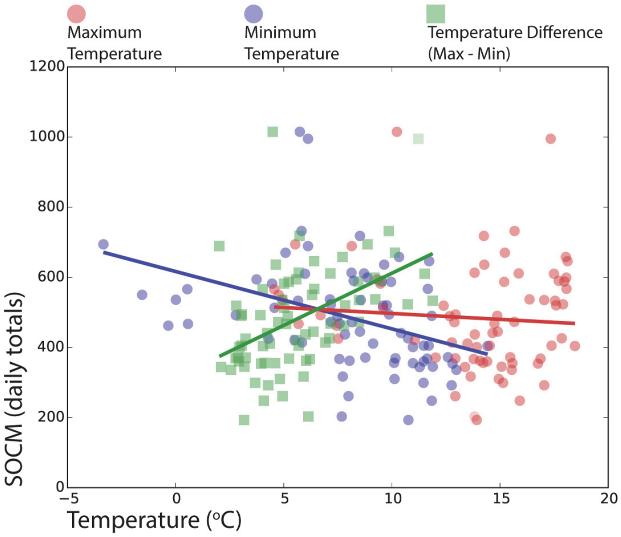

Figure 3: SOCM vs. Temperature

It is evident from the graph that no correlation exists between maximum temperatures and SOCM, at least for the time range in question, while daily minimum temperatures exhibit some degree of negative correlation with SOCM. It is interesting to note though that temperature difference between maximum and minimum recorded daily temperatures provides the best fit of the three variables from a statistical point of view, with a positive correlation. However, as temperature difference does not directly relate to an attribute that could explain this behaviour, analysis turns to other climate characteristics, more specifically cloud coverage, wind speed, and precipitation probability and intensity, which should at the same time affect SOCM as well as temperatures. These characteristics are known to affect ground temperatures (Easterling et al. 1997), and can be considered as creating unfavourable conditions for park visitors, thus reducing visitor numbers.

Cloud coverage exhibits a negative correlation with SOCM, with a strong (for the dataset) fit, as seen on Figure 4. Similar results are displayed when comparing SOCM against wind speed, indicating that unfavourable weather conditions have a negative impact on park usage, as would be expected. SOCM and precipitation exhibit a similar relationship, although not linearly correlated. As seen on Figure 5, for precipitation values greater than 0 (chance of precipitation), SOCM values average at about 400 daily total posts, providing a potential baseline of park activity regardless of weather conditions, possibly indicating restaurant visitors and less weather-dependent activities, such as exercise activities.

Hourly Aggregate

Analysis at a daily resolution identified some weather characteristics as broad drivers of park visitor activity, as shown previously. In this next section, activity will be investigated at an hourly temporal resolution, in order to capture the rela-

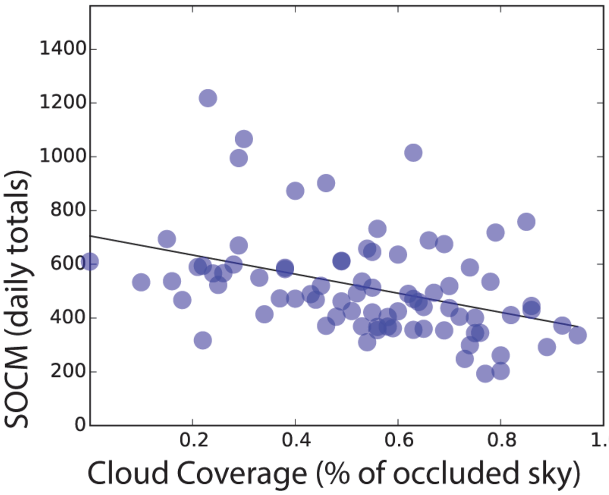

Figure 4: SOCM vs. Cloud Coverage

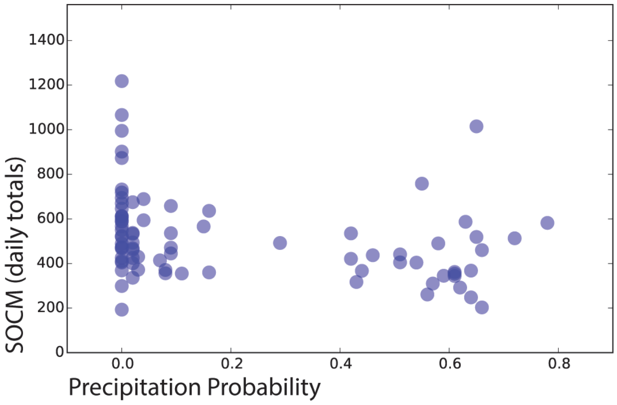

Figure 5: SOCM vs. Precipitation Probability

tionship between SOCM and weather/temporal characteristics in more detail. Looking at hourly SOCM totals against hourly temperature, as shown in Figure 6, it again becomes clear that on the whole, there exists no clear relationship between temperature and park visitor activity. Within individual days, there is a constant positive correlation between park activity and hourly temperature, as would be expected, given that higher temperatures coincide with daylight hours associated with high urban activity (afternoon hours).

Similar results of no correlation whatsoever are exhibited when looking at other weather characteristics at an hourly temporal scale, such as cloud coverage or wind speed. Data points in these cases are scattered with no discernible patterns, with the exception of precipitation intensity, where, as expected, SOCM values are at their constant lowest (approx. 20 per hour) when any rainfall is recorded. Of course, this behaviour of no relationship at hourly levels is expected. Given the temporal scale of one hour, variation in SOCM is caused more by hour of day and daily activity cycles, than any other climate characteristic. Following this reasoning, a very discernible pattern is exhibited when looking at SOCM by hour of the day, as seen in Figure 7.

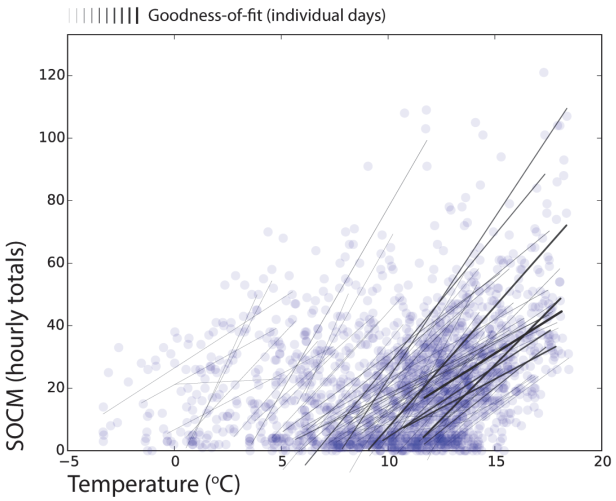

Figure 6: SOCM vs. Temperature - Hourly

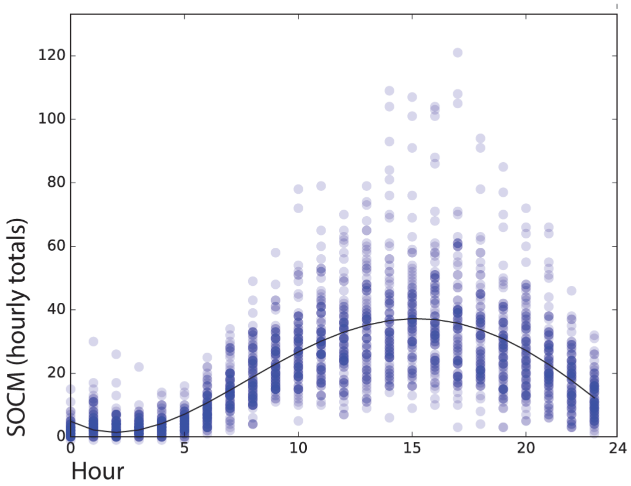

Figure 7: SOCM vs. Hour

Hourly SOCM values are at their lowest during early morning hours, between midnight and 5 am, with valley values at 2 am. Activity starts to pick up at 6 am, and rises steadily until a peak is reached at 3 pm. After this hour, values decrease again steadily into the night, until they are at their lowest at 2 am again. This oscillation in SOCM values can be modelled using a 4th degree polynomial, in the form of $y = ax^4 + bx^3 + cx^2 + dx + e$, with $a = 0.001, b = -0.065, c = 1.15, d = -3.8, e = 4.87$, which when fitted to the data points, results in a coefficient of determination of 0.47.

Variance in hourly SOCM values can be further explained as a result of weather effects at this point, as hour of day and curve fitting can provide a baseline 'default' behaviour, or in other words, average visitors based only on time of day, with all other conditions being equal. By additionally plotting point size and colour as a function of hourly precipitation probability, as seen on Figure 8, it is evident that points with higher precipitation probability fall below the curve. In other words, for a given hour, the number of park visitors can be predicted first by the hour of the day, and secondly by weather conditions. Very similar results are exhibited when substituting precipitation probability with other observed weather conditions, such as cloud coverage and wind speed, as the independent variable.

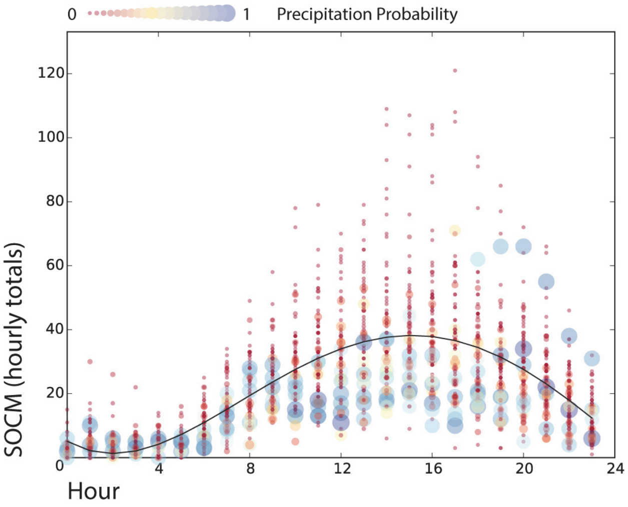

Figure 8: SOCM vs. Hour-Precipitation - Hourly

Validation

This work used social media posts as a proxy of actual park visitors, in an attempt to analyze the effect of weather and temporal parameters on park visitor numbers. In order to extrapolate back to real visitor numbers, short site surveys were conducted on four occasions within 2 weeks, where park visitor numbers were captured using manual counts. On each occasion the same route was taken, covering as much area as possible, visitors within 100 meters of the route were counted, and each survey lasted approximately 95 minutes. Following the survey, all social media posts originating from the area of interest within the duration of the survey were captured. One site survey was discarded, as a popular event was taking place at that day, making capturing visitor numbers impossible.

On inspection, social media posts appear to follow actual park visitor numbers, at a ratio of approximately 47 people per SOCM, as seen in Figure 9, providing some initial credibility to the use of SOCM as a proxy for actual activity. Some further validation and verification of assumptions was further gained from the surveys. More specifically, surveys 1 and 2 took place on a weekday, while survey 3 took place on a Sunday, same as the discarded survey, illustrating the increased park visitor numbers on Sundays. Furthermore, over the course of survey 1, weather started to change, from a clear sky to overcast and rainy, while on the other two occasions the sky was clear with sporadic clouds, indicating that unfavourable weather characteristics have a negative effect on park visitor numbers. Therefore, these small observations hint further at a valid hypothesis, although the sample is still far too small to conclude with any degree of certainty.

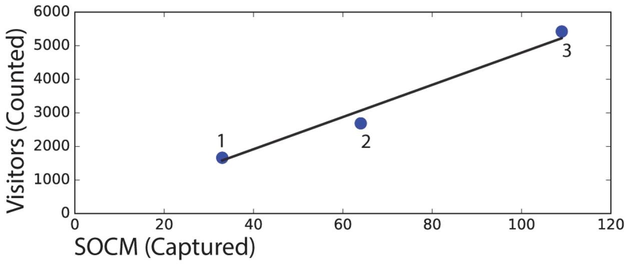

Figure 9: Visitors vs. SOCM

Conclusion and Further Work

This paper presented a correlation between park visitor activity, and weather and temporal characteristics. It was shown that park visitor activity is primarily driven by the time of day, resulting in predictable daily life cycles for the park, with weather conditions having a secondary effect. Furthermore, this study found some correlation between park visitor activity and social media posts originating from the area of interest, which although based on a small sample, hints at some constant value of 1 social media record for every 47 actual visitors.

Although this paper presented an analysis of archived data, it is interesting to note that these records were captured at the moment of their generation, i.e. in real time. Given the findings of this study, correlations between time of day/weather and SOCM could then be used in real time, by employing weather forecasts for the near-future and other real-time data sources to estimate current activity, and subsequently verify the result very soon afterwards, continuously. Such an approach could provide a perpetual model of public space activity, with continuous prediction and verification.

The findings in this study aim to enhance Public Space Use Studies, as discussed previously, by providing an opportunity to capture data remotely. Given the requirements and costs of conducting full-scale surveys, and thus their infrequent use, this method discussed here can provide data for undocumented periods, in order to 'fill in the gaps'. Additionally, given the fast turnover of results (real-time), this approach to public space visitor estimation can offer a quick monitoring platform for interested parties, for example park authorities.

This paper presented findings of an on-going study. It will continue to record and archive data, in order to cover larger periods of time, as well as to capture more active periods of the year, i.e. summer months. Furthermore, additional studies will be conducted in other public spaces, in order to build a more comprehensive archive of public space use activity. Additionally, the data and methods presented in this study will be combined with Agent-Based Simulations of park activity, where data generated from the above methods will be used as input in interaction models, to be used for vizualisation and further spatial analysis of park visitor activity and interaction.

Acknowledgements

This research was funded in part by Intel ICRI Cities. Manual surveys were conducted using the Android application 'Fieldworker', developed by Panos Mavros and Katerina Skroumpelou at CASA, UCL (github fork: github.com/pmavros/Fieldworker).

Notes

The datasets used in this study are available from http://cheliotk.info/various/CL2016-RT-Data.zip.

References

Batty, M.; Axhausen, K. W.; Giannotti, F.; Pozdnoukhov, A.; Bazzani, A.; Wachowicz, M.; Ouzounis, G.; and Portugali, Y. 2012. Smart cities of the future. *The European Physical Journal Special Topics* 214(1):481–518.

Calabrese, F.; Colonna, M.; Lovisolo, P.; Parata, D.; and Ratti, C. 2011. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems* 12(1):141–151.

Carmona, M. 2010a. Contemporary Public Space: Critique and Classification, Part One: Critique. *Journal of Urban Design* 15(1):123–148.

Carmona, M. 2010b. Contemporary Public Space, Part Two: Classification. *Journal of Urban Design* 15(2):157–173.

Diao, M.; Zhu, Y.; Ferreira, J.; and Ratti, C. 2015. Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design* 0265813515600896.

Dines, N. T.; Catell, V.; and Gesler, W. M. 2006. *Public Spaces, Social Relations and Well-being in East London*. Bristol, UK: Policy Press.

Easterling, D. R.; Horton, B.; Jones, P. D.; Peterson, T. C.; Karl, T. R.; Parker, D. E.; Salinger, M. J.; Razuvayev, V.; Plummer, N.; Jamason, P.; and Folland, C. K. 1997. Maximum and Minimum Temperature Trends for the Globe. *Science* 277(5324):364–367.

Gehl, J., and Gemze, L. 2000. *New city spaces*. Danish Architectural Press.

Gehl, J. 1987. *Life between buildings: using public space*. New York: Van Nostrand Reinhold.

Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; and Ratti, C. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41(3):260–271.

ICRI. 2015. *ICRI Annual Report 2013-2014*. Intel Cities.

Krier, R. 1979. *Urban space*. Rizzoli International Publications.

Ratti, C., and Claudel, M. 2014. LIVE Singapore! The Urban Data Collider. *Transfers* 4(3):117–121.

Stefanidis, A.; Crooks, A.; and Radzikowski, J. 2011. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78(2):319–338.

Townsend, A. M. 2000. Life in the Real-Time City: Mobile Telephones and Urban Metabolism. *Journal of Urban Technology* 7(2):85–104.

Whyte, W. H. 1980. *The Social Life of Small Urban Spaces*. Conservation Foundation.

Whyte, W. H. 1988. *City: Rediscovering the Center*. Anchor Books.