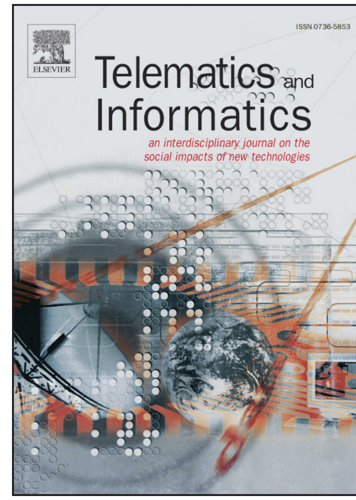

Urban Data and Urban Design: A Data Mining Approach to Architecture Education

Francesc Valls, Ernesto Redondo, David Fonseca, Ricardo Torres-Kompen, Sergi Villagrasa, Nuria Martí

PII: S0736-5853(17)30341-6
DOI: https://doi.org/10.1016/j.tele.2017.09.015
Reference: TELE 1010

To appear in: *Telematics and Informatics*

Received Date: 30 May 2017
Revised Date: 7 July 2017
Accepted Date: 28 September 2017

Please cite this article as: Valls, F., Redondo, E., Fonseca, D., Torres-Kompen, R., Villagrasa, S., Martí, N., Urban Data and Urban Design: A Data Mining Approach to Architecture Education, *Telematics and Informatics* (2017), doi: https://doi.org/10.1016/j.tele.2017.09.015

Urban Data and Urban Design: A Data Mining Approach to Architecture Education

Francesc Valls
francesc.valls@upc.edu
UPC (Polytechnic University of Catalonia), Barcelona Tech.
Superior Technical School of Architecture of Barcelona.
Avda. Diagonal, 649, 2. - ZIP code: 08028, Barcelona. Spain.

Ernesto Redondo
ernesto.redondo@upc.edu
UPC (Polytechnic University of Catalonia), Barcelona Tech.
Superior Technical School of Architecture of Barcelona.
Avda. Diagonal, 649, 2. - ZIP code: 08028, Barcelona. Spain.

[Corresponding Author]
David Fonseca
fonsi@salle.url.edu
La Salle, Ramon Llull University
GRETEL- Research Group in Enhanced TEchnology Learning
C/ Sant Joan de la Salle, 42 - ZIP code: 08022, Barcelona, Spain.
Phone number: (+34) 93 290 24 72 - Fax number: (+34) 93 290 24 20

Ricardo Torres-Kompen
ricardot@salleurl.edu
La Salle, Ramon Llull University
GRETEL- Research Group in Enhanced TEchnology Learning
C/ Sant Joan de la Salle, 42 - ZIP code: 08022, Barcelona, Spain.

Sergi Villagrasa
sergiv@salle.url.edu
La Salle, Ramon Llull University
GRETEL- Research Group in Enhanced TEchnology Learning
C/ Sant Joan de la Salle, 42 - ZIP code: 08022, Barcelona, Spain.

Nuria Martí
nmarti@salle.url.edu
La Salle, Ramon Llull University
Architecture Department
C/ Sant Joan de la Salle, 42 - ZIP code: 08022, Barcelona, Spain.

Abstract

The configuration of urban projects using Information and Communication Technologies is an essential aspect in the education of future architects. Students must know the technologies that will facilitate their academic and professional development, as well as anticipating the needs of the citizens and the requirements of their designs. In this paper, a data mining approach was used to outline the strategic requirements for an urban design project in an architecture course using a Project-Based Learning strategy. Informal data related to an award-winning public space (Gillett Square in London, UK) was retrieved from two social networks (Flickr and Twitter), and from its official website. The analysis focused on semantic, temporal and spatial patterns, aspects generally overlooked in traditional approaches. Text-mining techniques were used to relate semantic and temporal data, focusing on seasonal and weekly (work-leisure) cycles, and the geographic patterns were extracted both from geotagged pictures and by geocoding user locations. The results showed that it is possible to obtain and extract valuable data and information in order to determine the different uses and architectural requirements of an urban space, but such data and information can be challenging to retrieve, structure, analyze and visualize. The main goal of the paper is to outline a strategy and present a visualization of the results, in a way designed to be attractive and informative for both students and professionals –even without a technical background– so the conducted analysis may be reproducible in other urban data contexts.

Keywords: Mobile indoor content; Student mobile usability; User behavior; Virtual reality; Architectural education.

Abstract

The configuration of urban projects using Information and Communication Technologies is an essential aspect in the education of future architects. Students must know the technologies that will facilitate their
academic and professional development, as well as anticipating the needs of the citizens and the requirements of their designs. In this paper, a data mining approach was used to outline the strategic requirements for an urban design project in an architecture course using a Project-Based Learning strategy. Informal data related to an award-winning public space (Gillett Square in London, UK) was retrieved from two social networks (Flickr and Twitter), and from its official website. The analysis focused on semantic, temporal and spatial patterns, aspects generally overlooked in traditional approaches. Text-mining techniques were used to relate semantic and temporal data, focusing on seasonal and weekly (work-leisure) cycles, and the geographic patterns were extracted both from geotagged pictures and by geocoding user locations. The results showed that it is possible to obtain and extract

valuable data and information in order to determine the different uses and architectural requirements of an urban space, but such data and information can be challenging to retrieve, structure, analyze and visualize. The main goal of the paper is to outline a strategy and present a visualization of the results, in a way designed to be attractive and informative for both students and professionals –even without a technical background– so the conducted analysis may be reproducible in other urban data contexts.

Keywords: Data Mining, Urban Data, Architecture Education, Informal Learning.

1 Introduction

According to the Royal Institute of British Architects (RIBA) in its Plan of Work 2013[1] (Sinclair, 2013), the first key stage in a building project is "Strategic Definition", where the core project requirements are identified. In this stage, it is crucial to identify the requirements that need to be fulfilled by the proposed architectural or urban design.

Architectural education has traditionally relied on Project-Based Learning (PBL), where students are required to develop a proposal, usually over the course a semester, in a process that mimics the workflow of an architectural studio. During the development of this proposal, students learn to integrate often-conflicting aesthetic, constructive, structural, environmental, and usability requirements into a cohesive design, under the guidance of a tutor. In this scheme, the students are usually provided with the location where the design is to be developed and examples of related notable designs as reference.

Architects and urban designers (both graduate and undergraduate) learn about their discipline in a continuous and informal way, because the subject of their craft surrounds them almost anywhere and anytime, thus explaining the important historic role of travel in the formative years of architects. However, nowadays the world that surrounds us is increasingly digital, especially for the younger generations using mobile devices and cloud computing services (Moreira & Ferreira, 2017; Moreira, Ferreira, Pereira, & Durao, 2016), and in the specific framework of architectural education and professional practice it is clear that we should incorporate this new paradigm and approaches.

From the criticism of the mechanistic approach to urban planning in the decade of 1960 (Jacobs, 1961; Alexander, 1965), a tradition of the study of the city from the point of view of its users has a long tradition that can be traced from Kevin Lynch (Lynch, 1960) to Jan Gehl (Gehl & Svarre, 2013).

[1] https://www.ribaplanofwork.com/

However, despite the enormous amount of urban data, the architectural and urban design fields are yet to incorporate many sources of information into their workflow.

Representation technologies are used throughout the architectural design process to bring ideas into reality, allowing communication between designers, clients, contractors and collaborators (Horne and Thompson, 2008). Architecture students must learn to be proficient in these representation technologies throughout their studies, and must reach the point where drawing and representation blend together, and drawing becomes thinking (Suwa & Tversky, 1997). Therefore, it is necessary that students become skillful in multiple representation technologies, and that they are capable of incorporating the latest technologies into their design process in order to better communicate their proposals, and to facilitate critical reasoning on the spaces they conceive.

Following previous research about public participation (Fonseca, Valls, Redondo, & Villagrasa, 2016) and the feasibility of extracting information from Cadastral data (Valls, Garcia-Almirall, Redondo, & Fonseca, 2014), this paper discusses the process of extracting knowledge from informal online sources (Russell, 2014), to provide educational materials to architectural students in order to define the project requirements for an architectural design.

With this objective in mind, informal data related to a public space (Gillett Square in London, UK) was extracted from two social networks (Flickr and Twitter) and from its official website. The analysis was focused on semantic, temporal and spatial patterns, aspects generally overlooked in traditional approaches, to improve the education of future urban designers and architects in order to relate the projects to the main needs of the citizenship. This type of analysis had to be a replicable example for the students in order to implement an ongoing project of re-urbanization of public spaces in Barcelona, where the students were participating. Using the implemented method, students and future architects should be able to incorporate informal data obtained from citizens in order to improve their capabilities and digital skills in the representation of information. Thus, designs could be executed with a suitable design and adapted to the space in addition to combining the functionality, the needs and the interests of citizens.

In section 2, the framework related to the technological innovations in the urban and architecture studies is contextualized. Section 3 includes the explanation of the data retrieval and the transformation that followed. Finally, the main results of the study are presented and discussed in section 4, leading to the conclusions in section 5.

2 Framework

2.1 ICTs in Urban and Architecture Data Visualization and Analysis

Information and Communication Technologies (ICTs) are transforming citizens' lifestyles, adding new dimensions to the concept of socialization, as well as creating new habits (Oulasvirta, Rattenbury, Ma, & Raita, 2012). Other studies (Bower, Cram, & Groom, 2010) describe the opportunities offered by these emerging technologies as "creating a new kind of reality, one in which physical and digital environments, media and interactions are woven together throughout our daily lives." At the same time, new university students can be defined as Digital Natives (Margaryan, Littlejohn, & Vojt, 2011) or Digital Residents (White & Cornu, 2011), because they coexist and use all kinds of network technologies, multiple applications and all kinds of mobile devices at very early ages.

Until recently, in architectural education, the use of ICT was restricted to project implementation processes, where various applications such as Computer Assisted Design (CAD) and Building Information Modeling (BIM) served merely as aids in the execution of one's work (Navarro & Fonseca, 2017). Historically, in civil and building engineering education, visualization and understanding of 3D space was typically accomplished via the classical view (physical models and drawings), in front of 3D models and using virtual specifications. This approach is changing due to a generational change and the continuous improvement and development of technology. The new systems based on Virtual and Augmented Reality (VR/AR), Geo-Referencing, and learning gamification, will gradually reduce the control imposed on the designed tasks and scheduled presentations. Due to the potential of virtual systems, the spatial skills and abilities of students can be reinforced, while also using the essential interactive and collaborative features of these processes. Students can work with peers and teachers and participate in multi-tasking/multi-user collaborative and instant tracking (Calongne, 2008).

Focused on urban data, it was proposed (Gordillo, Gallego, Barra, & Quemada, 2013) a generic model to support a new way of visiting a city. In this approach, instead of understanding the city as a place for tourism, the students perceive it as a place for learning in which all necessary educational resources are available. The model has been conceived as a way to encourage learners to create their own educational tours, in which Learning Points of Interest are set up to be discovered using two models: formal (conducted by a teacher) and informal outdoor mobile learning (where no educator is directly related to the learning experience).

Recently, there has been an increasing interest in Learning Analytics (LA) in Technology Enhanced Learning (TEL). The TEL research field has been deeply involved with the development and application of collaboration apps. TEL seeks to improve the students' learning experience by supporting student engagement, satisfaction and retention; helping to produce enterprising graduates with the skills required to compete in the global business environment; encouraging inspirational and innovative teaching; personalizing learning that promotes reflection; and delivering and supporting continuing professional development and internationalization (Vicent, Villagrasa, Fonseca, & Redondo, 2015).

LA is also a field in which several related areas of research in TEL converge. These include academic analytics, action research, educational data mining, recommender systems, and personalized adaptive learning (Chatti, Dyckhoff, Schroeder, & Thüs, 2012). Big data (BD) or Data Mining (DM) applied to education are emerging interdisciplinary research fields also known as educational data mining (EDM) (Romero & Ventura, 2013). It is concerned with developing methods for exploring the unique types of data that come from educational environments (Boyd & Crawford, 2012). Its goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain educational phenomena (Monika Goyal, 2012).

2.2 Case of Study

The elective multimedia course in the Barcelona School of Architecture introduces the students to emerging technologies such as augmented and virtual reality, photogrammetry and 3D printing. In the academic years 2015-2016 and 2016-2017, the course focused on using videogame technology for architecture representation (Valls, Redondo, & Fonseca, 2015; Valls, Redondo, Fonseca, Garcia-Almirall, & Subirós, 2016), taking advantage of improvements in real-time rendering to produce interactive content. During the 2016-2017 edition, the students participated in an educational experience placed at the intersection of architectural representation and urban design.

Following the constructivist approach (Papert, 1980) in urban planning (Hjorth & Wilensky, 2014), a proposed urban renovation project was used as a case of study for the duration of the course. The proposal consisted in the conversion of some streets in Barcelona into pedestrian-only streets, and the creation of civic squares at their intersections. The course was split into roughly two parts: at the beginning of the course, the students were divided into groups of three to four students, and each group was assigned to work on part of the urban environment. During the following weeks, the groups modeled and textured the façades of their respective sections, following simple guidelines regarding aspects such as the polygon count or the size of the textures of their models, since they would later be used as assets in the game engine. At the end of this process, all the models were consolidated into a single environment, shared by all the groups.

In the second part of the course, the students worked with the game development platform, which in this case was Unreal Engine (Epic Games, 2012). In this stage the students were asked to produce assets based on the proposals made by the students of the Master in Landscape Architecture, who designed a multi-purpose module/kiosk (Figure 1) measuring 1x2.3m, and 2m high, to be used by the community in the newly created public spaces.

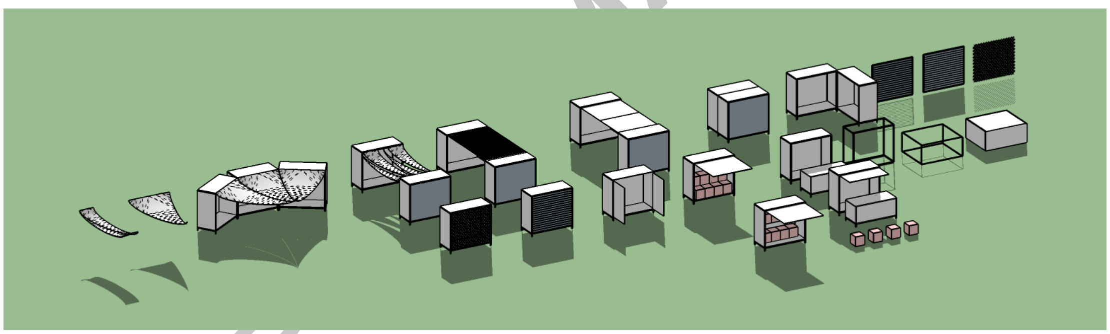

Figure 1: Modeled variations of the module in the early stages of development

The students had to adapt the kiosk design to their own proposals of possible uses, and model and texture them accordingly. In the course, they also learnt to define the necessary programming to move around the simulated environment responding to user inputs, and grab any of their modules and change their location or rotate them with the mouse wheel (Figure 2).

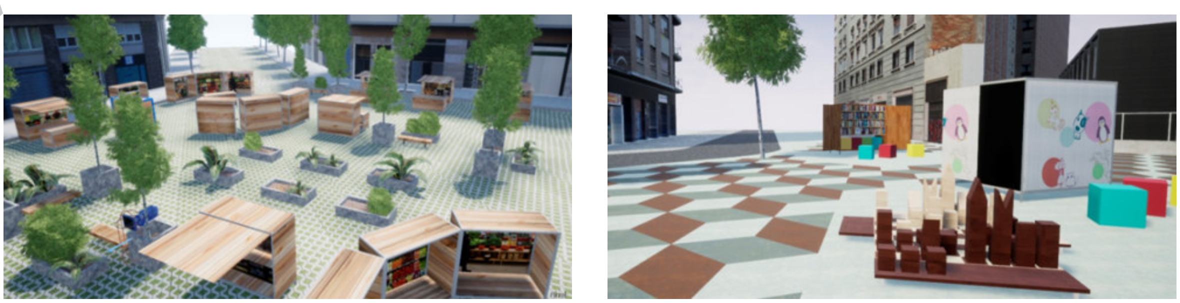

Figure 2: Some of the students' proposals, inserted into the simulated environment and with the capacity to be interacted with (moved and rotated)

2.3 Urban Data and Design Proposals

During the course, stakeholders suggested that a possible reference for the students of a similar civic space could be Gillett Square, located in London (UK), a community-led regeneration project in a former car park, designed by Hawkins-Brown Architects, and recipient of a WAN Award in 2012 in the Effectiveness category[2]. The proposal to analyze the project carried out in Gillett Square, and its possible implications on the global project that will be held in Barcelona, was suggested by the municipality of Barcelona, one of the entities that is giving explicit support to the completion of the overall project. Since the students could be considered digital residents and found social media content engaging (Shen, Liu, & Wang, 2013; Tur, Marín-Juarros, & Carpenter, 2017), it was decided to retrieve, analyze and visualize content from Twitter using text mining tools. The results were uploaded to the educational intranet and served as starting point for the development of the students' proposals.

In addition, geo-tagged pictures from Flickr and the content of the official Gillett Square website[3] regarding the schedule and content of past events organized in the square were also retrieved and analyzed, with the objective of being used as teaching materials in the next edition of the course.

3 Data Retrieval and Transformation

The research used publicly available data sourced from the Gillett Square website and from its official Twitter account. This officially sourced data was complemented with user-generated content from the Flickr photo sharing community (Table 1).

Table 1: Summary of data sources

Source	Start Date	End Date	Retrieval Date	Retrieval Method	Format	Records	Characters	Status
Flickr	01/01/2010	31/12/2016	14/03/2017	Flickr web API	JSON	35502	1002879	Unofficial
Twitter	06/07/2012	30/03/2017	31/03/2017	Twitter web API	JSON	1105	131721	Official
Calendar	16/09/2014	21/04/2017	13/04/2017	HTTP GET	HTML	190	128183	Official
Website	18/06/2012	25/07/2016	13/04/2017	HTTP GET	iCal	104	89846	Official

The data retrieval process was through the corresponding web Application Programming Interfaces (API) to request the desired data from Flickr and Twitter, or directly requested from the website. The obtained data formats (JSON, HTML and iCal) had to be parsed to be converted into data frames suitable for analysis. The time spans for the different data sources overlapped partially except during the last quarter

[2] https://backstage.worldarchitecturenews.com/wanawards/award/sector/effectiveness-12
[3] http://www.gillettsquare.org.uk/home

of 2014, the whole year 2015 and the first half of 2016, where data was available from all 4 sources (Figure 3).

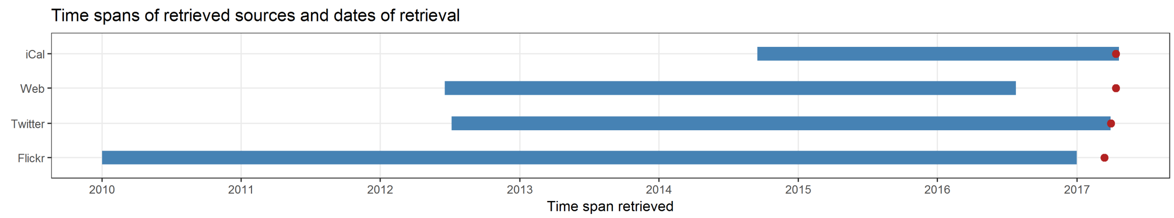

Figure 3: Time spans of the retrieved sources (as blue bars from the earliest recorded date to the latest) and corresponding dates of retrieval (as red dots), showing periods of overlapping temporal data

The reason for the choice of sources was threefold:

- Included both official and user-generated content to explore the suitability of using unofficial data where no official data is available
- Allowed comparing the content of the Twitter timeline to more structured data from the website, complementing the more informal nature of the content sourced from social media, and help counteract some of the population bias of its users' demographics (Longley, Adnan, & Lansley, 2015)
- Obtained data both through APIs and directly from the web, and therefore sourced live dynamic data from content providers as well as from the open (crawlable) web

3.1 Retrieving Geotagged Pictures through the Flickr API

The Flickr API allows requesting pictures inside a specific geographical area defined by bounding rectangle through the coordinates of its southwest and northeast corners. To retrieve the geotagged pictures around the area of interest, a geofence consisting on a 0.03 x 0.03 degree area centered on the center of the area of interest was defined. The API required authentication using an API Key, which is transmitted as part of the request URL through the HTTPS protocol. The data was retrieved using the R package curl 2.6 (Ooms, 2017), and 35,502 unique records were obtained in JavaScript Object Notation (JSON) format between the requested dates of January 1st 2010 and December 31st 2016.

The JSON strings were parsed using the R package jsonlite 1.4 (Ooms, 2014) and their latitude and longitude coordinates were extracted, along with other variables. These latitude and longitude pairs

Coordinate Reference System (CRS) was EPSG:4326[4] (WGS 84), used by GPS devices, and were projected onto EPSG:3857[5] (WGS 84 / Pseudo Mercator), used by web mapping applications.

3.2 Retrieving Twitter Data

The home page of the Gillett Square website linked to the official Twitter account @gillettsquare[6] on Twitter. At the time of writing, Twitter had become an important source of news and social networking, with users posting, reading and replying through short messages (less than 140 characters) using mobile devices or web browsers.

Twitter users have the capacity of following a specific account and the entire list of publicly posted messages of a user (timeline) is available for browsing. In addition, users can prepend certain words with a "#" symbol (hashtag), post URLs (usually shortened due to space constraints) and embed or link multimedia content (e.g. pictures or videos).

Twitter allows access to some of their data through its API (with some limitations), and requires authentication using OAUTH. There are several software components that allow interacting with the Twitter API using different programming languages; for this research the R package rtweet 0.4.0 (Kearney, 2016) was used, which can interface with the stream and REST APIs and convert the JSON responses into data frames. Other R packages were also considered: twitteR 1.1.9 (Gentry, 2015) and streamR 0.2.1 (Barbera, 2014), as well as the Python library Tweepy[7].

Using the Twitter API, the complete public timeline (list of status updates or "tweets") was retrieved, along with additional metadata, obtaining 1105 messages from July 6th 2012 to March 30th 2017. The list of followers (users subscribed to the account) was also retrieved, obtaining a list of 1134 users.

This list of users was further processed, requesting information about these profiles through the API. After looking up their profiles, 1115 returned data (98%), obtaining the profile metadata (e.g. description, language, location, UTC offset and time zone).

[4] https://epsg.io/4326/
[5] https://epsg.io/3857
[6] https://twitter.com/gillettsquare
[7] https://www.tweepy.org/

3.3 Retrieving Calendar Data

The Gillett Square website provided a browseable list of events[8] that linked to the specific page that explained each event in detail; this list contained 10 events and at the bottom of this list of events there were a set of buttons to navigate to other pages with information about the 10 next or previous events, and a button to download the information shown on the page as an iCal file.

Unfortunately, analyzing the HTML structure of the page, it was discovered that the link was generated by JavaScript code embedded in the page that dynamically altered the page structure and generated the link in connection with a cloud service that stored the calendar, and therefore it was not possible to parse the raw HTML code to obtain the link[9].

However, the page had a search box at the top of the page that built a query as a URL that requested the data to the server, and it was possible to emulate this search behavior using the R package httr 1.2.1 (Wickham, 2016a) and build the necessary URL strings to retrieve the pages with the desired data, obtaining 19 iCal files that contained 10 events each, from September 16th 2014 to April 21st 2017.

The resulting files were parsed according to RFC5545, Internet Calendaring and Scheduling Core Object Specification for iCalendar (Desruisseaux, 2009). The format was human-readable plain text and the information was extracted using Regular Expressions with the R packages stringr 1.2.0 (Wickham, 2016b) and stringi 1.1.5 (Gagolewski, 2017).

The text stored in the “DESCRIPTION” property of each event was the main body of the text and could consist on multiple paragraphs; to be able to process its content with the text analysis tools, newlines and tabs were converted to spaces, escaped commas were converted to commas, and extra blank space was trimmed. The title of the event was retrieved from the “SUMMARY” property. The beginning and end each event was stored in the “DTSTART” and “DTEND” properties, where the date and time were combined in a single string, and were extracted using capturing groups for time and date[10]; once extracted the two parts, separators were introduced for dates[11] and times[12]. After parsing all 10 files, the results were combined into a data frame with 190 records (one for each event) with 6 fields each: summary, description, start date, start time, end date and end time.

---

[8] http://www.gillettsquare.org.uk/events/

[9] It was possible, however, running a headless browser using PhantomJS (http://phantomjs.org/)

[10] With the pattern ^DTSTART.*:(\d*)T*(\d*)$ for start, and ^DTEND.*:(\d*)T*(\d*)$ for end

[11] The pattern (\d{4})(\d{2})(\d{2}) was converted to \1-\2-\3

[12] The pattern (..)(?=.) was converted to \1:

3.4 Website Scraping

The Gillett Square website also had access to events older than the ones covered by the retrieved iCal data, reachable through the website sitemap[13] that gave access to an archived version of the website[14] where it was possible retrieve a page with a list[15] with links to all the events from June 18th 2012 to July 25th 2016, whose HTML content was retrieved. The R package rvest 0.3.2 (Wickham, 2016), which uses a similar philosophy as the Python library Beautiful Soup[16] was used to scrape information from this page, examining its structure with the Selector Gadget[17] bookmarklet to identify the nodes that contained the desired information with a combination of XPATH and CSS selectors. This allowed the retrieval of three pieces of information for each of the 104 events:

- Its title
- The date on which it took place
- A link to a page describing the event in detail

After retargeting these links to the archived version of the site, the raw HTML code of all the linked pages was retrieved and parsed, extracting the text describing the details of the events, again using XPATH and CSS to isolate de desired nodes in the Document Object Model (DOM) tree of the pages.

3.5 Data Transformation

After parsing the data sources to extract the desired information into a table format, the data had to be transformed to be suitable for analysis. The manipulation followed the principles of Tidy Data (Wickham, 2011, 2014), where each variable is stored in its own column and each observation is saved in its own row, and reproducible research (Peng, 2011), and therefore all data transformation operations were performed through code from the original sources.

After the retrieval and parsing stages, the amount of text collected was significant but was not immediately suitable for analysis, as it consisted on an unstructured collection of long strings of characters. The R packages tm 0.7-1 (Feinerer, Hornik, & Meyer, 2008) and quanteda 0.9.9-50 (Benoit et al., 2017) were used to cleanup and tokenize the text to obtain a document-term matrix (DTM) suitable for text analysis with natural language processing tools. The Python libraries Natural Language Toolkit

---

[13] http://www.gillettsquare.org.uk/sitemap
[14] http://www.gillettsquare.org.uk/events.bak
[15] http://www.gillettsquare.org.uk/events.bak?filter=past&pp=all
[16] https://www.crummy.com/software/BeautifulSoup/
[17] http://selectorgadget.com/

3.2.2 (Bird, Klein, & Loper, 2009) and scikit-learn 0.18 (Pedregosa et al., 2011) were also considered, but were not used because it was preferred stay within the R environment.

The string manipulation operations used the R package stringr 1.2.0 (Wickham, 2016) based on the package stringi 1.1.5 (Gagolewski, 2017), which in turn is based on the International Components for Unicode (ICU) libraries for Unicode compliance.

The text to be analyzed was sourced from online sources and contained some elements different than what text mining tools expect as input. The texts were pre-processed to discard these elements:

- Non-alphanumeric or non-punctuation characters (e.g. tabs and newlines) were removed and converted to white space
- All characters were converted to lowercase using the Unicode guidelines
- HTML escape characters (“&”, “<” and “>”) were converted to Unicode (“&”, “<” and “>”)
- Link URLs were removed
- Twitter hashtags (#) were removed
- Twitter screen names (@) were removed

The pre-processed texts were converted into a corpus, an object that stores and indexes the texts (documents) along with their corresponding document-level variables (docvars) as well as document-level and corpus-level metadata. In the text analysis stage, the document-level variables would be used to subset the corpus according to specific properties, and aggregate the corpus into groups sharing the same variable to explore similarities between them.

Once stored in a corpus, the documents were tokenized (Table 2), breaking the document character streams into meaningful elements (tokens). The first step of the tokenization was the removal of certain features:

- Ordinals, usually days of the month (e.g. 1st, 12nd)
- Times, generally in the 12h format (e.g. 11am, 5pm)
- Prices (numbers, including the decimal point, prepended by a currency sign)
- Numbers
- Punctuation

- In some analysis, specific words such as "gillett" and "square"
- Extra white space trimming
- Compound words were separated (e.g. "two-day" became "two" and "day")

Table 2: Number of tokens from each of the retrieved sources, before and after transformations

Source	Original		Transformed	
	Tokens	Unique tokens (types)	Tokens	Unique tokens (types)
Twitter	26558	23049	6966	1411
Website	16599	10640	7434	1576
Calendar	23104	15830	10392	1252

In addition, common English stop words (which provide little meaning but appear frequently in English sentences) were removed, using the SMART[18] list (Salton, 1971) with the addition of the word "will" (which do not appear in the list). Since Twitter message length limitations favors word economy, a thesaurus was built to consolidate common weekday and month names with their abbreviations into a single token, and avoid counting them separately (e.g. "feb" and "February" or "wed" and "Wednesday"). After this process, the tokens were stemmed using the Porter stemming algorithm (Porter, 1980), using the R package SnowballC 0.5.1 (Bouchet-Valat, 2014), to reduce the inflected terms to their root form. A more complex approach using lemmatization with TreeTagger[19] was not deemed necessary for the purposes of the present study, because the text was not syntactically complex and the focus was on the word frequency.

As the last step, the stemming process was reversed using stem completion (Feinerer, 2010), assigning the most frequent word in each group of words that had been reduced to the same stem, to display a meaningful word in the analyses as a proxy of the stem. In addition, since the Porter stemmer alters the last character in many words ending in "y" (e.g. "july" becomes "juli"), words that did not find a match in the stem completion process and ended in "i" were corrected accordingly. At the end of the process, the documents were converted into a Document-Term Matrix (DTM), sometimes called Document-Feature Matrix[20] (DFM), used by natural language processing software, generally weighted using the term frequency - inverse document frequency (tf-idf) statistic. In these matrices, the rows correspond to each

[18] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[19] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[20] Since it can contain features other than terms, such as n-grams, skipgrams or multi-word expressions

document in the corpus, and the columns to each term, while at the intersection of each document and term appears the count of each term in a specific document.

Since any document contains only a subset of the words, the matrix is generally sparse, with many terms (columns) containing only zeros; to reduce the size and complexity of working with the matrix, the least frequent words can be dropped, reducing the sparsity of the matrix.

The DFM matrix also contains all the document variables present in the corpus, making possible to refactor the matrix into the groups defined by these variables.

3.6 Temporal Data

Because of the complexity of working with temporal data, which has many irregularities (different text representations in human and machine language, months with uneven number of days, leap years, leap seconds, time zones, daylight saving time, weekly cycles, non-decimal units of measurement as hours, minutes or seconds, etc.), all date and time conversion and manipulation was performed with the R package lubridate 1.6.0 (Grolemund & Wickham, 2011), which simplifies working with date-times and time-spans.

All retrieved information had temporal data associated but had to be converted from its text representation to a date-time object in the POSIXct format (which stores the number of seconds since the beginning of 1970 (in the UTC time zone).

The package was used to extract the month component of de date, which in turn determined the season according to the criteria of the UK Met Office ("Weather and climate change," 2016), which in addition to the astronomical seasons defines the following meteorological seasons:

- Spring[21]: March, April and May
- Summer[22]: June, July and August
- Autumn[23]: September, October and November
- Winter[24]: December, January and March

[21] http://www.metoffice.gov.uk/learning/learn-about-the-weather/how-weather-works/seasons/spring
[22] http://www.metoffice.gov.uk/learning/learn-about-the-weather/how-weather-works/seasons/summer
[23] http://www.metoffice.gov.uk/learning/learn-about-the-weather/how-weather-works/seasons/autumn
[24] http://www.metoffice.gov.uk/learning/learn-about-the-weather/how-weather-works/seasons/winter

The day of the week was also extracted from the dates, to separate work days (Monday, Tuesday, Wednesday, Thursday and Friday) from weekends (Saturday and Sunday), and explore the differences in the type of events organized in days categorized as work or leisure through their textual description. Finally, the start and end dates of each event was used to build a matrix with as many rows as analyzed events (190) and as many columns as minutes in a day (1440). For each event (row), the matrix stored a zero if there was no event scheduled in the corresponding minute (column) and a one if there was. This one-minute binning was chosen because it was the maximum resolution of the available schedule.

This matrix was transformed with the R package reshape2 (Wickham, 2007) from an array representation (190 rows x 1440 columns) to a long form containing 273600 rows with three variables: the row index, the column index and the value. This matrix was then joined to the corresponding event data (by row) to incorporate the other variables (day of the week, season).

3.7 Location Data

The data from the followers of the @gillettsquare Twitter account included a text string with the location of these users. This location is filled out voluntarily by the users when they register to the service, and oftentimes include witty or humorous remarks; however, although it is not completely accurate it can contain valuable data about the location of the users.

The data was geocoded using the Google Geocoding API[25], building a geocoding request URL with the appropriate query parameters, and obtaining a JSON response. The used query parameters included region biasing to favor results in London, using the UK country code top-level domain (ccTLD), which is not identical to the ISO 3166-1 code (GB). The resulting JSON was parsed to obtain the latitude/longitude coordinate pairs and the accuracy of the result (e.g. country, state, city, street); in the cases the service returned an array of multiple results, the first one was chosen.

In addition to the Google geocoder, the Data Science Toolkit[26] geocoder was also queried using the R package ggmap 2.6.1 (Kahle & Wickham, 2013). This geocoder emulates the JSON response of the Google service, but does not have its license limitations. Comparing the number of successfully geocoded locations of the 1115 users, the Google service returned 893 (80%), while the Data Science Toolkit returned 871 (78%).

[25] https://developers.google.com/maps/documentation/geocoding/
[26] http://www.datasciencetoolkit.org/

4 Results

Following the data retrieval and transformation processes, the data was analyzed and visualized. The conducted analyses were carefully selected with two main objectives:

- Allow the teaching staff to define the project program for the course based on the extracted knowledge, focusing on aspects that are usually overlooked because of the lack of available data
- Generate visually engaging teaching materials to support the students' creative process when designing their proposals

4.1 Semantic Content

The rich textual data retrieved allowed the exploration of the content using word clouds, which encode the relative frequency of the word as the size of the text, producing chart that is space-efficient to display and compare the word frequencies in a body of text, and can be both visually engaging and easy to interpret.

In addition, two variations of the word cloud chart were used to explore the influence on the text content of two factors (seasons in the yearly cycle, and workdays and weekends in the weekly cycle): comparison clouds and commonality clouds (Fellows, 2014).

A comparison cloud compares the frequencies of words across factors, placing each factor at equal intervals around a circle. The size of each word is proportional to its frequency within its group (and therefore sizes are not comparable across them), and more frequent words are placed at the center. Words are colored according to the sectors defined by the factors, and words at the center axis of the sector are more strongly related to their factor, with the strength decreasing when moving off-axis to another factor.

As a complement to the comparison clouds, commonality clouds plot the cloud of words shared across the factors compared.

The analysis was based on the retrieved calendar data (subsection 3.3), complemented by data from the archived website (subsection 3.4) for dates prior to 16th September 2014, except the data delivered to the students in the pilot course, which was based on data retrieved from Twitter (subsection 3.2).

4.1.1 Twitter Data in the Pilot Course

In the pilot course, the students were provided two word clouds, accessible through the educational intranet: (1) a static hashtag cloud of the Twitter @gillettsquare account, and (2) a dynamic clickable word cloud with the most frequent words contained in the tweets (Figure 4).

The first word cloud (Figure 4, left) was produced using the R package wordcloud 2.5 (Fellows, 2014), with the colors assigned according to a weighted quantile (where the weight was the hashtag frequency) to emphasize the most frequent hashtags. The interactive word cloud (Figure 4, right) was produced with the R package wordcloud2 0.2.0 (Lang, 2016) based on wordcloud2.js[27], which dynamically generates a web page that renders the words using HTML5 technology, allowing the user to hover on any word to overlay its frequency.

Figure 4: Static hashtag cloud (left) and interactive word cloud (right) provided to the students in the pilot course

4.1.2 Seasonal Variation in the Yearly Cycle

When designing an open-air public space, students and professionals must contemplate the seasonal variations throughout the year, which influence climate factors[28] (such as temperature, weather, and available natural light during of the day) or social life (e.g. festivities such as New Year celebrations or school holidays).

Text corpus content was grouped in four groups corresponding to the four seasons defined by the UK MetOffice, according to the date on which the event took place, and a comparison plot (Figure 5, left) was produced placing each season in a quadrant:

[27] https://github.com/timdream/wordcloud2.js
[28] At latitudes far from the equator

- Winter mentioned more strongly the word “ice”, and occasional outdoor activities (“market”, “choir”), as well as countries with cold climates (“Russia”, “Norway”)
- Summer[29] was more strongly associated with outdoor activities (“dance”, “music”, “performance”, “majorettes”), and with words related to holidays (“school”, “workshop”)
- Spring and autumn were less defined, mentioning outdoor activities tied to a specific theme (“film”, “shakespeare”, “carnival”)

The corresponding commonality plot (Figure 5, right) showed the words shared across the four seasons, most of which were related to children and families and/or playful activities, suggesting that throughout the year and regardless of the season, family-oriented events are planned.

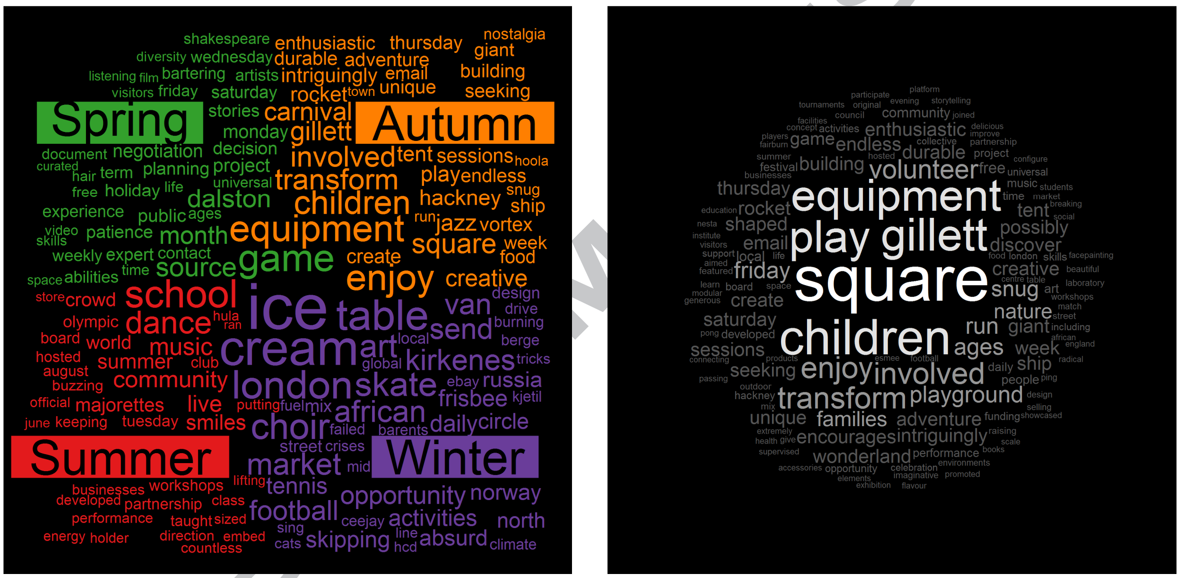

Figure 5: Comparison cloud (left) and commonality cloud (right) illustrating seasonal variations in the text content

4.1.3 Work-Leisure Variation in the weekly Cycle

Beyond the seasonal variation, which is a natural consequence of the yearly cycle, it exists an artificial weekly cycle[30] with an alternating pattern of work and leisure days, in which the use patterns of public space can vary dramatically. To find out the different activities conducted in this work-leisure axis, the text corpus was partition into two groups according to the day of the week the described event took place: Monday-Friday (work) and Saturday-Sunday (leisure).

[29] Interestingly, “olympic” appeared in the Spring-Summer boundary because the London 2012 Summer Olympics

[30] Which roughly corresponds to one fourth of a moon cycle

Despite the ratio between work days and weekends being 5:2, the resulting comparison cloud (Figure 6, left) showed that the majority of the high frequency words are linked to work, while leisure is associated with a large variety of low frequency words:

- Surprisingly, "children" is associated with the work pole, which suggests that the events involving children are more geared towards after school activities rather than weekend activities
- The word "Friday" is at the frontier of both poles, arguably because of the ambivalence of Friday afternoon
- The leisure activities are more varied, and as expected feature words related to leisure ("music", "dance", "performance", "food", "live"," art", "festival", "market", etc.)

The corresponding commonality plot (Figure 6, right) showed the words shared across the two poles, exhibiting a very similar pattern to the seasonal commonality plot discussed above, reinforcing its conclusions.

Figure 6: Comparison cloud (left) and commonality cloud (right) illustrating variations in the text content across weekdays (work) and weekends (leisure)

4.2 Event Schedule

Knowledge about temporal use patterns of public spaces is crucial in urban planning and urban design education and professional practice, to generate proposals suitable to the interests of its users. However, this information is scarce and difficult to obtain and analyze. For this study, the event schedule retrieved from the calendar data was used, as it included the start and end times of each event over the course of almost three years.

Following the semantic analysis discussed in subsection 4.1.3, the temporal patterns along three different cycles (Figure 7) were analyzed from the event schedule data retrieved from the calendar, and a suitable visual representation was chosen accordingly:

- Daily cycle (event distribution along the length in minutes of the day)
- Weekly cycle (distribution along the days of the week)
- Yearly cycle (distribution along the months and seasons of the year)

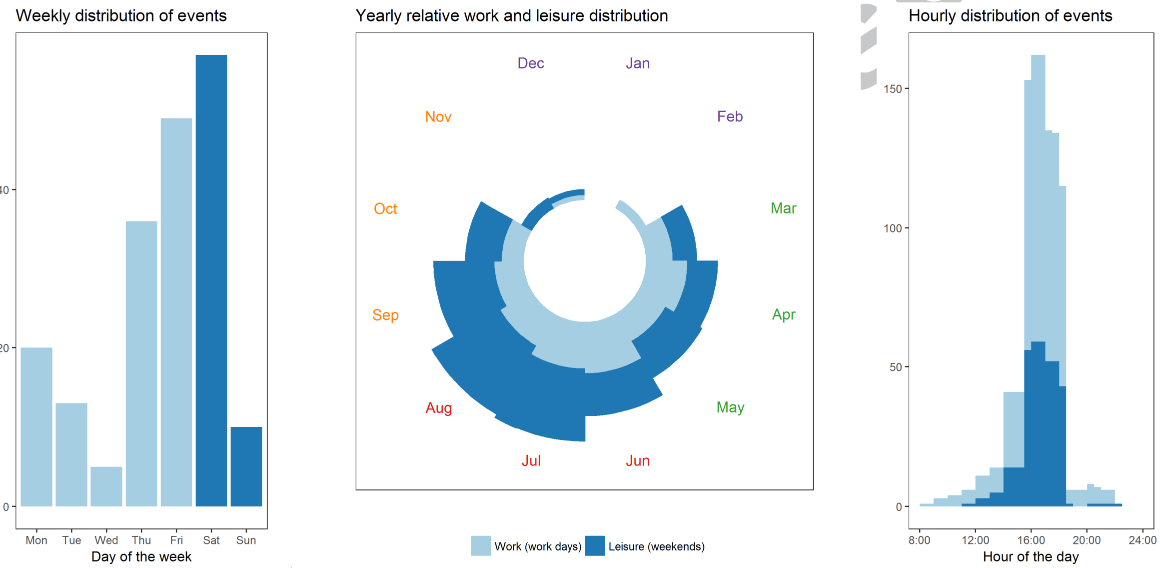

Figure 7: Weekly distribution of events (left), relative yearly distribution of events (center), and daily distribution of events (right), colored according to work days (Monday to Friday) and weekends (Saturday and Sunday), with the same color scheme as the comparison cloud in Figure 6

The daily cycle (Figure 7, right), was visualized with a histogram with the distribution of minutes during the day when an event was taking place, excluding the minutes in the 0-8AM interval, where no events took place. The distribution was stepped because most of the events began and ended at the top of the hour or in multiples of 30 minutes. The histogram shows that most of the events took place in the 3:30PM-6:30PM interval, and that activities in weekends are specially concentrated in this interval, while on work days the distribution is more spread before and after.

The weekly cycle (Figure 7, left) counted the number of events, taking place on each day of the week, sorted from Monday to Sunday according to the ISO 8601 standard[31]. Each event was counted once,

[31] https://www.iso.org/iso-8601-date-and-time-format.html

regardless of its duration. The distribution had its lowest on Wednesdays, and from Thursday to Saturdays increased steadily, to decrease sharply on Sundays. The 30% of the events took place on Saturday, and 75% of them on the Thursday-Friday-Saturday period.

The yearly cycle (Figure 7, center), was represented as the relative work and leisure distribution of the events per month in a circular diagram, where events taking place on weekends had a weight of 1/2, and events taking place on work days had a weight of 1/5, to take into account the probability of an event to happen according to the different sizes of both groups. The proportion of both events was roughly constant throughout the year, but the summer months (especially July and August) hosted more events while in winter the activity decreased significantly, with no events recorded on January.

As discussed in subsection 4.1.2, seasonal variations have a significant incidence in the use patterns of public space, that students must take into account in their proposals. This influence was explored in more detail using the event schedule retrieved from the calendar data (Figure 8).

The distribution of events along the year (Figure 8, left) showed that Summer is the season with more scheduled events, followed by Spring and Autumn, while in Winter there were almost no events scheduled.

The hourly distribution of events (Figure 8, right) displayed the same trend and reflected the changing duration of the days across seasons, with the hourly distribution of events beginning earlier and finishing later in summer, to become less spread out as the days get shorter in Spring and Autumn.

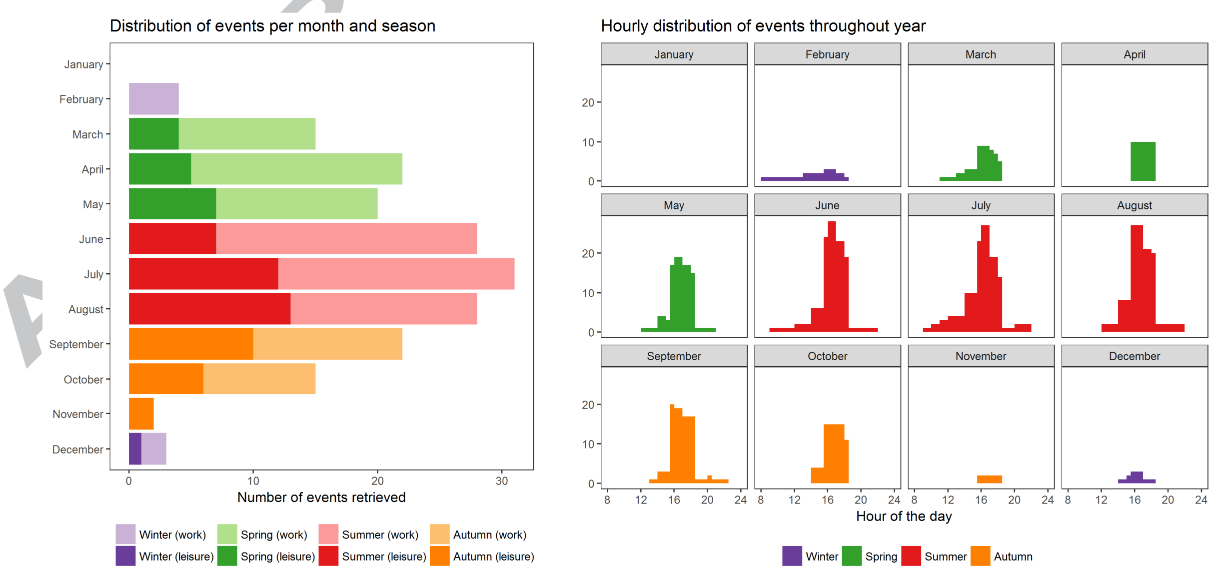

Figure 8: Distribution of events per month (left) and variation of the hourly distribution of events along the year (right). Both charts use the same color scheme applied to seasons as the comparison cloud in Figure 5

4.3 Spatial Patterns

In the fields of architecture and urban planning and design, the context where the proposal is to be inserted (its physical landscape, but also its cultural background) is a central element, and students are required to develop a strong sense of place. However, the globalization of the professional practice, where architecture firms participate in international competitions, requires making proposals in places architects may not familiar with.

The data mining approach used to analyze the semantic and temporal content was also used to explore the area of influence of the public space of the case of study in three administrative boundaries: city, borough and district (Figure 9). The results showed that the smaller geographic areas were mentioned more frequently, suggesting that the geographic scope of this particular public space was local; these results also matched the data in the hashtag cloud (Figure 4) discussed in subsection 4.1.1.

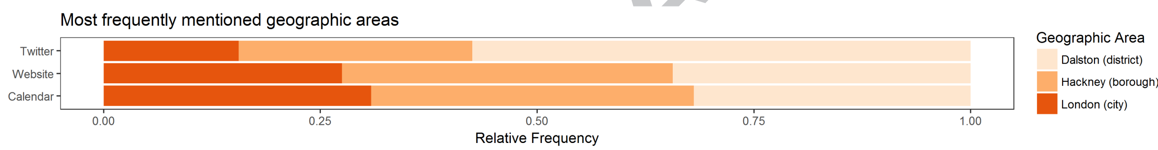

Figure 9: Scope of geographic areas mentioned in the three sources used, from largest to smallest: city (London), borough (Hackney) and district (Dalston)

In architectural education, the site map is the document where the intentions of the intended design are made explicit. The possibility of producing information-rich site maps using data retrieved from social media was explored using the retrieved data from Flickr and Twitter (Figure 10).

Using the digital geo-located data of the 35502 photos retrieved from Flickr overlaid on a base map produced a site map (Figure 10, left) depicting the most photographed sites around the area of interest as a proxy of pedestrian behavior, where the main walking axis were revealed as well as the most attractive landmarks. The majority of the pictures follow the main roads, except in the south of the map, where the outline of Regent's Canal is clearly visible.

In addition, the locations of the successfully geocoded addresses of the followers of the Twitter account of the case of study were also plotted over a base map. The locations were mainly distributed around the area next to the square, confirming the findings discussed previously, but allowed a more nuanced interpretation, showing that the followers were located around the site but more concentrated towards the south, and that the river Thames acted as a barrier.

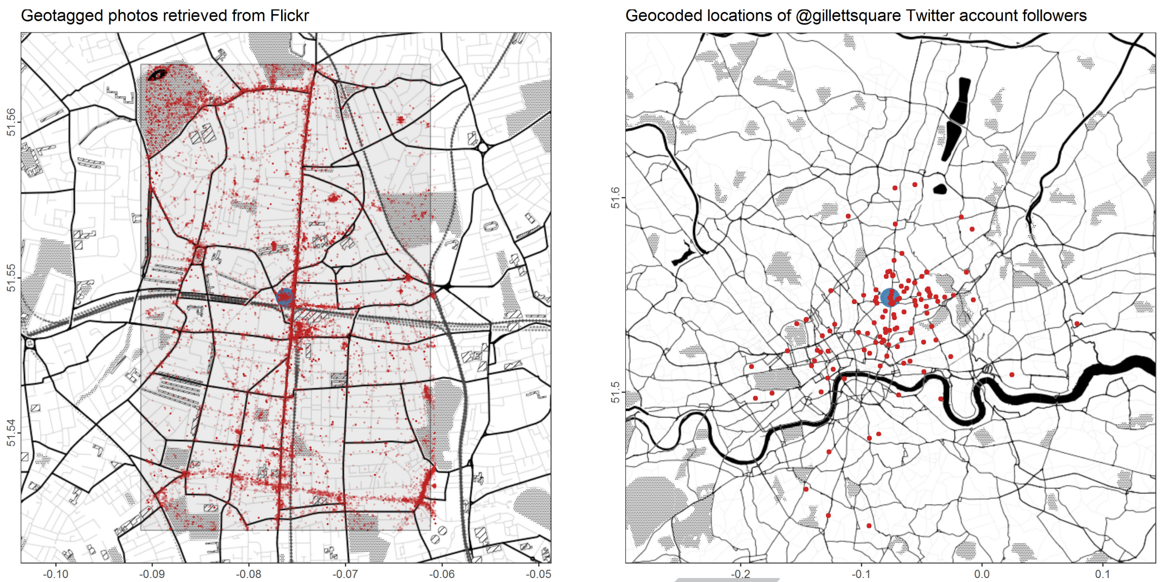

Figure 10: Locations of the geo-located pictures retrieved around the case of study (left) and geocoded addresses of the followers of the @gillettsquare Twitter account close to the case of study (right). Base maps by Stamen Design (http://stamen.com) from Open Street Map data

5 Conclusions and future work

In many disciplines, the boundary between education and professional practice is difficult to define, especially in the architectural and urban planning fields, where education has traditionally followed a PBL approach and students assume the role of a trained professional in a professional studio. In this context, this research explored online data, either stored in web pages or informally generated by users and posted on social media, as a source of information for urban planners and designers.

The authors focused on semantic, temporal and spatial patterns, which until the advent of online sources have lacked enough data to conduct exploratory research. The results showed that it is possible to extract very valuable information, but it can be difficult to retrieve, structure, analyze and visualize. However, the authors believe that the conducted research is reproducible in other urban data contexts, and that the visualization of the results is attractive and informative for professionals without a technical background. In addition to obtaining data for the initial definition of the project, social media data should become an additional tool to evaluate the citizens' response in public participation processes before the final design, and to gather informal feedback about its suitability after the completion of the project.

On the other hand, the increasing internationalization of architectural studios, where teams submit their proposals in international competitions, has increased the geographic scope of the architectural practice, challenging the architects' ability to adapt to multiple cultural backgrounds beyond what they are more familiar. In this context, online data could provide valuable information on the site in the early stages of the proposal.

The analysis of the spatial patterns for educational purposes will be explored in future editions of the elective subject on Geographic Information Systems in the Barcelona School of Architecture, and the engagement of students using data from social media will be measured in comparison with previous editions of the course. In additional other techniques of automated knowledge extraction will be applied to additional cases of studies (e.g. museums, universities, sports facilities, shopping centers) to validate suitability and improve the developed methodology. Following these proposals, we cannot forget the preparation of teacher in order to give the correct support to students (Moreira, Pereira, Durao, & Ferreira, 2017). This issue is critical in order to include mobile education and informal learning in the skills and curriculum of our students by conducting good technological practices.

The generic approach used to analyze the data allows generalizing the conclusions and applications to other educational fields. The informal data extraction and its uses can improve the digital skills and academic development of our students, independent of the framework. Analyzing social data, students can develop more sustainable projects and products adapted to more users and/or users with different profiles or disabilities.

The study presented in this paper has been developed in parallel with a pilot project to model new uses in a square of Barcelona. In this case, virtual reality and realistic rendering models were developed to configure collaborative and interactive uses for the city. The following stages (in the next academic course) will replicate and improve the study presented in this paper. For this proposal, both the current state of the urban zones to by re-defined and the proposals developed during the last academic course will be studied in two different courses of the schools participating in the project. In this case, it will be the first opportunity to compare and discuss the results of both cities and the main concepts that urban designers and architecture students must be able to take into account in order to optimize the uses of urban spaces.

Acknowledgments

This research was supported by the National Program of Research, Development and Innovation aimed to the Society Challenges with the references BIA2016-77464-C2-1-R & BIA2016-77464-C2-2-R, both of the National Plan for Scientific Research, Development and Technological Innovation 2013-2016, Government of Spain, titled "Gamificación para la enseñanza del diseño urbano y la integración en ella de la participación ciudadana (ArchGAME4CITY)", & "Diseño Gamificado de visualización 3D con sistemas de realidad virtual para el estudio de la mejora de competencias motivacionales, sociales y espaciales del usuario (EduGAME4CITY)". (AEI/FEDER, UE).

Bibliography

Alexander, C. (1965). A City is not a Tree. *Architectural Forum*, *122*(1), 58–62.

Barbera, P. (2014). *streamR: Access to Twitter Streaming API via R*. Retrieved from https://CRAN.R-project.org/package=streamR

Benoit, K., Watanabe, K., Nulty, P., Obeng, A., Wang, H., Lauderdale, B., & Lowe, W. (2017). *quanteda: Quantitative Analysis of Textual Data*. Retrieved from http://quanteda.io

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1st ed). Beijing ; Cambridge [Mass.]: O'Reilly. Retrieved from http://www.nltk.org/book/

Bouchet-Valat, M. (2014). *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library*. Retrieved from https://CRAN.R-project.org/package=SnowballC

Bower, M., Cram, a, & Groom, D. (2010). Blended reality: Issues and potentials in combining virtual worlds and face-to-face classes. *ASCILITE 2010 - The Australasian Society for Computers in Learning in Tertiary Education*, 129–140.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Calongne, C. M. (2008). Educational Frontiers: Learning in a VIRTUAL WORLD. *Educause Review*, *43*, 36–48. https://doi.org/10.1007/s00371-007-0182-z

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning*, *4*(5/6), 318–331. https://doi.org/DOI: 10.1504/IJTEL.2012.051815

Desruisseaux, B. (2009, September). Internet Calendaring and Scheduling Core Object Specification (iCalendar). Internet Engineering Task Force. Retrieved from https://tools.ietf.org/html/rfc5545

Epic Games. (2012). *Unreal Engine 4*. Epic Games.

Feinerer, I. (2010). Analysis and Algorithms for Stemming Inversion. In P.-J. Cheng, M.-Y. Kan, W. Lam, & P. Nakov (Eds.), *Information Retrieval Technology* (Vol. 6458, pp. 290–299). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17187-1_28

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5). https://doi.org/10.18637/jss.v025.i05

Fellows, I. (2014). *wordcloud: Word Clouds*. Retrieved from https://CRAN.R-project.org/package=wordcloud

Fonseca, D., Valls, F., Redondo, E., & Villagrasa, S. (2016). Informal interactions in 3D education: Citizenship participation and assessment of virtual urban proposals. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2015.05.032

Gagolewski, M. (2017). *R package stringi: Character string processing facilities*. Retrieved from http://www.gagolewski.com/software/stringi/

Gehl, J., & Svarre, B. (2013). *How to Study Public Life* (2nd edition). Washington, DC: Island Press.

Gentry, J. (2015). *twitteR: R Based Twitter Client*. Retrieved from https://CRAN.R-project.org/package=twitteR

Gordillo, A., Gallego, D., Barra, E., & Quemada, J. (2013). The city as a learning gamified platform. In *Proceedings - Frontiers in Education Conference, FIE*. https://doi.org/10.1109/FIE.2013.6684850

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3). https://doi.org/10.18637/jss.v040.i03

Hjorth, A., & Wilensky, U. (2014). Redesigning Your City – A Constructionist Environment for Urban Planning Education. *Informatics in Education*, *13*(2), 197–208. https://doi.org/10.15388/infedu.2014.02

Horne, Margaret and Thompson, E. M. (2008). The role of virtual reality in built environment education. *Journal for Education in the Built Environmen*, *3*, 5--24.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. New York: Random House.

Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, *5*(1), 144–161.

Kearney, M. W. (2016). *rtweet: Collecting Twitter Data*. Retrieved from https://cran.r-project.org/package=rtweet

Lang, D. (2016). *wordcloud2: Create Word Cloud by htmlWidget*. Retrieved from https://CRAN.R-project.org/package=wordcloud2

Longley, P. A., Adnan, M., & Lansley, G. (2015). The Geotemporal Demographics of Twitter Usage. *Environment and Planning A*, *47*(2), 465–484. https://doi.org/10.1068/a130122p

Lynch, K. (1960). *The Image of the City*. Cambridge (MA): The MIT Press.

Margaryan, A., Littlejohn, A., & Vojt, G. (2011a). Are digital natives a myth or reality? University students' use of digital technologies. *Computers & Education*, *56*(2), 429–440. https://doi.org/10.1016/j.compedu.2010.09.004

Margaryan, A., Littlejohn, A., & Vojt, G. (2011b). Are digital natives a myth or reality? University students' use of digital technologies. *Computers & Education*, *56*(2), 429–440. https://doi.org/10.1016/j.compedu.2010.09.004

Met Office. (2016, June 3). Weather and climate change. Retrieved May 12, 2017, from http://www.metoffice.gov.uk/

Monika Goyal, R. V. (2012). Applications of Data Mining in Higher Education. *IJCSI International Journal of Computer Science Issues*, *9*(2), 113–120.

Moreira, F., & Ferreira, M. J. (2017). No TitleTeaching and Learning Requirement Engineering based on Mobile Devices and Cloud: A Case Study. In *Blended Learning: Concepts, Methodologies,* *Tools, and Applications* (pp. 1190–1217). Hershey: Information Science Reference. https://doi.org/10.4018/978-1-5225-0783-3.ch059

Moreira, F., Ferreira, M. J., Pereira, C. S., & Durao, N. (2016). Evolution and use of mobile devices in higher education: A case study in Portuguese Higher Education Institutions between 2009/2010 and 2014/2015. *Telematics and Informatics*. https://doi.org/10.1016/j.tele.2016.08.010

Moreira, F., Pereira, C. S., Durao, N., & Ferreira, M. J. (2017). Mobile Learning in Portuguese Universities: Are Professors Ready?. Recent Advances in Information Systems and Technologies. *Advances in Intelligent Systems and Computing*, *570*, 887–898. https://doi.org/10.1007/978-3-319-56538-5_88

Navarro, I., & Fonseca, D. (2017). Nuevas Tecnologías de Visualización para mejorar la representación de arquitectura en la educación. *Architecture, City, Environment - ACE*, *in edition*.

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]*. Retrieved from https://arxiv.org/abs/1403.2805

Ooms, J. (2017). *curl: A Modern and Flexible Web Client for R*. Retrieved from https://CRAN.R-project.org/package=curl

Oulasvirta, A., Rattenbury, T., Ma, L., & Raita, E. (2012). Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, *16*(1), 105–114. https://doi.org/10.1007/s00779-011-0412-2

Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. New York, NY, USA: Basic Books, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226. https://doi.org/10.1126/science.1213847

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Russell, M. A. (2014). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More* (2. ed). Beijing: O'Reilly.

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Shen, C.-X., Liu, R.-D., & Wang, D. (2013). Why are children attracted to the Internet? The role of need satisfaction perceived online and perceived in daily real life. *Computers in Human Behavior*, *29*(1), 185–192. https://doi.org/10.1016/j.chb.2012.08.004

Sinclair, D. (2013). *Plan of Work 2013*. London (UK): RIBA. Retrieved from https://www.ribaplanofwork.com/

Suwa, M., & Tversky, B. (1997). What Do Architects and Students Perceive in their Design Sketches? A Protocol Analysis. *Design Studies*, *18*(4), 385–403. https://doi.org/http://dx.doi.org/10.1016/S0142-694X(97)00008-2

Tur, G., Marín-Juarros, V., & Carpenter, J. (2017). Using Twitter in Higher Education in Spain and the USA. *Comunicar*, *25*(51). https://doi.org/10.3916/C51-2017-02

Valls, F., Garcia-Almirall, P., Redondo, E., & Fonseca, D. (2014). From Raw Data to Meaningful Information: A Representational Approach to Cadastral Databases in Relation to Urban Planning. *Future Internet*, *6*(4), 612–639. https://doi.org/10.3390/fi6040612

Valls, F., Redondo, E., & Fonseca, D. (2015). E-Learning and Serious Games: New Trends in Architectural and Urban Design Education. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies* (Vol. 9192, pp. 632–643). Los Angeles, CA, USA: Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-20609-7_59

Valls, F., Redondo, E., Fonseca, D., Garcia-Almirall, P., & Subirós, J. (2016). Videogame Technology in Architecture Education. In M. Kurosu (Ed.), *Human-Computer Interaction. Novel User Experiences* (pp. 436–447). Toronto, Canada: Springer International Publishing. https://doi.org/10.1007/978-3-319-39513-5_41

Vicent, L., Villagrasa, S., Fonseca, D., & Redondo, E. (2015). Virtual learning scenarios for qualitative assessment in higher education 3D arts. *Journal of Universal Computer Science*, *21*(8), 1086–1105.

White, D. S., & Cornu, A. L. (2011). Visitors and Residents: A new typology for online engagement. *First Monday*, *16*(9). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/3171

Wickham, H. (2007). Reshaping Data with the {reshape} Package. *Journal of Statistical Software*, *21*(12), 1–20.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23.

Wickham, H. (2016a). *httr: Tools for Working with URLs and HTTP*. Retrieved from https://CRAN.R-project.org/package=httr

Wickham, H. (2016b). *rvest: Easily Harvest (Scrape) Web Pages*. Retrieved from https://CRAN.R-project.org/package=rvest

Wickham, H. (2016c). *stringr: Simple, Consistent Wrappers for Common String Operations*. Retrieved from https://CRAN.R-project.org/package=stringr

ACCEPTED MANUSCRIPT

Highlights are mandatory for this journal. They consist of a short collection of bullet points that convey the core findings of the article and should be submitted in a separate editable file in the online submission system. Please use 'Highlights' in the file name and include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).

- Data mining for urban design projects in Architecture education.
- Informal and citizenship data in a Project-Based learning.
- Useful social network data to design new urban spaces.
- Semantic, temporal and spatial patterns analysis.
- Mobile technology uses and advanced data visualization.