Geographical distribution of the sources of IEQ complaints: an analysis through text-mining of online job reviews

Giorgia Chinazzo
Department of Civil and Environmental Engineering, Northwestern University, Evanston, USA

giorgia.chinazzo@northwestern.edu

Abstract. Text-mining allows analyzing a large amount of non-structured data, such as online reviews, to gain insights about previously unknown information. Online job reviews contain a variety of information, ranging from salary estimations to interview experiences. Among this information, the text posted online can report an evaluation of the workplace's indoor environmental quality (IEQ), describing both its positive and negative aspects. When referring to negative characteristics, online reviews can be considered to report IEQ complaints. Such complaints can be categorized according to the four IEQ aspects (i.e., thermal, visual, acoustic, and indoor air quality) and their combination. This paper exploits text-mining techniques to investigate the geographical distribution of the sources of IEQ complaints according to the location in which the job review is posted. The analysis is performed in terms of climate (according to the Köppen-Geiger climate classification), country, and population (to consider the distribution between high-density and low-density areas). The results show that the distribution of the source of IEQ complaints varies according to the climate and the country, even though thermal aspects are always the largest source of discomfort in all countries and climatic zones. The more significant rates of thermal complaints are observed in the U.S. and India. They could be associated with the extensive use of HVAC systems and the restrictive operating temperatures adopted in these countries. The results also show that acoustic, indoor air quality and visual complaints are more numerous in large cities than in rural areas, where thermal complaints prevail. This paper provides a picture of the current IEQ discomfort across several geographical regions and highlights the great potential of User-Generated-Content to study various aspects of the IEQ, in this case, their geographic distribution.

1. Introduction

The study of the indoor environmental quality (IEQ) is essential for improving the well-being and health of building occupants while operating buildings in an energy-efficient way. The collection of the occupants' feedback about the indoor environment is one of the most valuable tools to investigate and improve the IEQ. Conventionally, the occupants' feedback concerning existing and occupied buildings is collected in field studies through surveys [1]. Due to the limitations of surveys regarding cost and distribution style, efforts are being made to innovate how the occupants' feedback is collected in buildings. The most recent examples are the use of micro ecological momentary assessments on a smartwatch platform [2] and the continuous occupant voting systems (OVS) [3]. A completely new way to acquire occupants' feedback is through the scraping and analysis of information publicly posted online, which is defined as User-Generated-Content (UGC). UGC ranges from reviews to text, images,
and videos. The analysis of online reviews has been used in the hospitality sector to study the IEQ of hotels and Airbnb locations [4,5]. The website Glassdoor has recently been highlighted as a potential source of information about the IEQ of different workplaces worldwide [6].

This study is part of a larger project aiming to apply text-mining techniques to analyze UGC scraped from the Glassdoor website to study the IEQ of different workplaces [7]. The overarching goal of the project is to demonstrate the benefits of exploiting UGC for the study and improvement of the IEQ. Focusing on the negative reviews posted on the job website, this study aims at analyzing the information about the source of complaints referring to the IEQ of the workplace in combination with the geographical location of each review. Through a descriptive analysis of the unconventional dataset collected from the web, this study addresses the following research questions:

1) How does the distribution of complaints across the four IEQ aspects (i.e., thermal, visual, acoustic, indoor air quality – IAQ, and their combination) vary according to the climate?
2) How does the distribution of complaints across the four IEQ aspects vary according to the country?
3) How does the distribution of complaints across the four IEQ aspects vary between high-density and low-density areas?

2. Method
To address the research questions, this study exploits text mining techniques to extrapolate the feedback about the IEQ of different workplaces in various locations in the world from online job reviews. The method followed in this study is divided into three steps: (1) data selection, (2) data preparation, and (3) data analysis.

2.1. Data selection
The data selection phase includes the selection of the website from which to extrapolate the reviews, the organizations on which to focus the analysis, and the items to scrape from each review. This study focuses on the reviews scraped from Glassdoor, one of the largest and fastest-growing company review websites globally, containing reviews of more than 1 million companies from more than 190 countries [8]. Among all the reviews, only those of the companies with more than 10,000 employees are selected to focus the investigation on large organizations. The scraped reviews contain different kinds of information, ranging from the publication date to the employment status of the review's author. This study uses the information about the location of the review (and hence the job location) extracted from the author information section of the review (Figure 1(a)). The web-scrapping is executed through a custom script in R that allows obtaining the reviews of an organization's Glassdoor webpage once its web link is manually inserted in the script. Only English reviews are scraped from the website, and all the information is saved in a CSV file.

From the 1,158,706 job reviews scraped for this study, only 60.5% of them contain information about the location. This percentage refers to the number of reviews for which the "location" cell in the resulting CSV file is not empty or does not contain the expression "not given." However, it must be noted that some of these cells wrongly include other author information (i.e., mainly the job role). Considering the large number of reviews, it would be impossible (or overly time-consuming) to manually remove all the incorrect entries. For this reason, the geolocation of the scraped reviews illustrated in Figure 2 should only be intended as an indication of all the reviews' locations. The majority of the reviews indicating the location are reported in North America (72.9%), followed by Asia (15.9%) and Europe (8.5%). South America, Australia, and Africa cover the remaining 2.7%. Overall, reviews from 173 countries are collected, with most of the reviews located in the U.S. (68.8%), India (12.7%), the U.K. (4.2%), and Canada (4.0%). The uneven distribution of reviews in the world is due to the fact that only English reviews are analyzed.
2.2. Data preparation

The data preparation phase involves the extrapolation of IEQ complaints from the reviews and the geocoding of the location extracted from the author information section.

The method for the extrapolation of IEQ complaints is detailed by Chinazzo [7] and involves the preprocessing of the text extracted from the cons section (i.e., text transformation to lower case, and stop-words and punctuation removal) and the identification of words referring to one or more of the four IEQ aspects. This latter is performed through a so-called iterative cleaning with the use of a custom "IEQ vocabulary" containing IEQ words referring to the different IEQ aspects (e.g., warm and cold for the thermal aspect and dark and glary for the visual aspect) and of a custom "incorrect vocabulary" that allows excluding combinations of words not referring to the IEQ (e.g., "cold culture" or "make noise for yourself" are not categorized as thermal and acoustic complaints, respectively). The iterative cleaning also uses tokenization, the extrapolation of single words (i.e., unigrams) or two subsequent words (bigrams) from sentences, which are visually inspected and used to enrich the two vocabularies used in the iterative cleaning process. Once the reviews containing IEQ complaints are identified, they are categorized according to one of the IEQ aspects or their combination. Table 1 reports some examples of IEQ complaints according to the IEQ aspects.

The geocoding consists of identifying the latitude and longitude of each location reported in the reviews, usually indicated as "city, country or state." The geocoding process is performed using the "geocode" function of the R package ggmap [9] and Google Maps. Once the latitude and longitude are determined, each review is associated with a country and a climatic zone according to the Köppen-Geiger climate classification [10]. For each review, the country is determined with the function "coords2continent" of the R package rworldmap [11], and the climatic zone with the R package kgc [12]. Finally, each location is categorized according to the number of inhabitants: above 400,000 inhabitants, between 400,000 and 400,000 and below 40,000. Each city's population is extracted from the "world.cities" dataset of the R package maps [13], containing information of world cities of population greater than about 40,000. All the cities not reported in the dataset are categorized as having less than 40,000 inhabitants. Overall, of the 10,568 reviews containing IEQ complaints, only 6,903 reported a correct location, for a total of 3,098 unique locations. Due to the reduced number of reviews, a visual inspection allowed to remove from the final dataset the reviews indicating a location "not given" or wrongly reporting the job role.
Table 1. Examples of IEQ complaints by IEQ aspect.

| IEQ aspect | IEQ complaint |
|------------|--------------|
| Thermal    | "A/C system is dysfunctional all year long, and it has been the case for a long time. The office often gets freezing or boiling inside despite the temperature outside. Lots of people are getting sick because of it, but company will not take this into consideration when it comes to apply their sickness policy which used to be getting a warning after 3 instances of absence. Last summer you could see in the office very often agents having to work with blankets over their shoulders and coats during their shifts while it was 30 degrees outside" - Customer Service Guest Specialist in the U.K. |
| Acoustic   | "Distracting, loud environment. The open-desk style doesn't work for everyone (like me), I find it hard to concentrate and end up working at home more than I would like to, just to catch up." - Engineer in the U.S. |
| IAQ        | "Really old, dusty buildings with poor air quality. The enthusiasm to get started on workday dies as soon as I walk into the building. The air quality in building is starting to cause me health problems, I think I may develop lung cancer or die if I continue working here." - Product Engineer unknown location |
| Visual     | "Managers, senior I.T. consultants, all sit in the same I.T. factory cubes, the glare from outside light with the overhead florescence creates the perfect cocktail for cranium splitting headaches." - I.T. Consultant in the U.S. |
| IAQ & Acoustic | "My office was next to an assembly line, The noise was horrific. The place was also filthy dirty. We all had respiratory issues due to the conditions. Was also told by maintenance that only about 40% of the building's ventilation system was working. I went upstairs to escape these conditions only to find the dust was just as bad plus the carpets were dirty, moldy, and infested with vermin. I left after 11 months because I could not do my job in that environment." - Software Engineer in the U.S. |

2.3. Data analysis

Two analyses are performed on the preprocessed dataset to address the research questions of this study. First, the source of IEQ complaints by location (both climate and country) is calculated with the relative percentage of reviews containing IEQ complaints according to the IEQ aspects:

$$RP_{IEQ,k,i} = \frac{N_{IEQ,k,i}}{N_{IEQ,i}} \cdot 100$$

where \(N_{IEQ,k,i}\) is the number of reviews containing IEQ complaints referring to a specific IEQ aspect, \(k\), for a given climatic zone or country, \(i\), and \(N_{IEQ,i}\) is the number of reviews containing IEQ complaints in the same climatic zone or country. Such normalization is necessary considering the uneven distribution of reviews across climatic zones and countries (e.g., most reviews are reported in the USA). Then, the distribution of IEQ complaints per area population is calculated.

3. Results and discussion

3.1. Distribution of sources of IEQ complaints by climate

Figure 3(a) reports the distribution of complaints across the four IEQ aspects and their combination for the different climatic zones. The bars on the x-axis are ordered according to the number of IEQ complaints (from left to right, from the larger to the smaller number). Not all climatic zones are reported, but only those for which at least 20 IEQ complaints are present. The figure shows that the predominant source of complaint is thermal, independent of the climatic zone. The smallest relative percentage of thermal complaints is found in the Cfb and Csb climate zones, referring to an oceanic climate and a warm-summer Mediterranean climate, respectively. This reduced number of complaints could be associated with a milder climate (resulting in less extreme indoor conditions), better building construction or operation in the specific zones, or simply more extreme discomfort conditions associated
with the other IEQ aspects, especially the acoustic ones. The BSh (referring to a semi-arid climate) and the Cwa (referring to a humid subtropical climate) are the climatic zones reporting the highest rate of thermal complaints, but the different distribution could also be due to a reduced number of IEQ complaints in such climatic zones (74 and 33, respectively). Acoustic aspects appear as the second most reported source of complaint, mostly independently from the climatic zone. Only in the Dfb (humid continental climate) and Am (tropical monsoon climate) climatic zones, IAQ issues appear to prevail over the acoustic ones.

3.2. Distribution of sources of IEQ complaints by country

Figure 3(b) indicates the relative percentage of IEQ aspects per country. Also in this case, not all countries are reported, but only those for which at least 20 IEQ complaints are present. Visual and combined complaints are the less reported ones in all countries, except for visual complaints in Ireland, where they are associated with an almost double relative percentage compared to other countries. The figure shows that the larger rates of thermal complaints are observed in the U.S. and India. The higher thermal complaint rates could be associated with the extreme operating temperatures adopted in both countries due to the extensive use of HVAC systems, resulting in too cold environments in summer and too warm environments in winter. Such extreme buildings’ operation has an enormously detrimental impact on the energy consumption of buildings and the environment, other than negatively affecting the comfort of building occupants, and should be further addressed and regulated in these two countries. The figure also shows that the number of acoustic complaints is lower in the U.S. and Canada, and larger in Switzerland, probably due to the kind of workplaces in this country (supposed to be predominantly office buildings – see Chinazzo [7] for additional analyses on IAQ complaints and the type of workplace). The lack of control of job type in the analysis is a limitation of this study that could be further investigated in future research. IAQ complaints are more present in Canada, Australia, and The Netherlands and surprisingly less reported in India despite the renowned outdoor air pollution. The fact that people did not complain about the IAQ does not imply good filtration systems and acceptable IAQ. Still, it could indicate an alarming sign of adaptation to poor air quality conditions.

Figure 3. Relative percentage of IEQ aspects per climatic zone (a) and country (b). The x-axis is ordered according to the number of IEQ complaints in each group, from the largest to the smallest.
Figure 4. Geolocation of IEQ complaints in the world according to the type of IEQ aspects.

Figure 5. Geolocation of IEQ complaints in the U.S. according to the type of IEQ aspects.
Figure 4 illustrates the geolocation of the sources of IEQ complaints, with the size of the points representing the number of complaints in a specific location. The number and location of the points reflect the distribution of IEQ source of discomfort highlighted in Figure 3(b). For example, the mapped points in India show the reported problems about thermal and acoustic aspects and the limited presence of IAQ complaints. Also, the maps show the distribution of complaints in countries that are not reported in the previous figure, such as the presence of thermal and acoustic complaints in the east side of Asia and the lack of visual complaints in such areas. As most IEQ complaints are located in the U.S., Figure 5 illustrates a zoom into the country to further study the geolocation of the complaints. The figure shows that the larger hubs of complaints about acoustic, IAQ, and visual aspects are located in correspondence to big cities such as New York, Chicago, Atlanta, Austin, Seattle, San Francisco, and Los Angeles. It also shows that many hubs of complaints about thermal aspects are scattered around the country and are not only located in large cities. These observations inspired the analysis by population reported in the following subsection, in which the reviews from the U.S. are analyzed with those of the rest of the world.

3.3. Distribution of sources of IEQ complaints per area population

Figure 6 illustrates the relative percentages of IEQ complaints divided according to the IEQ aspects and the location population. The figure confirms what was previously observed from the maps of the U.S.: thermal complaints are predominant in more rural areas where smaller cities are present, whereas acoustic, IAQ, and visual complaints are more and more reported in densely populated areas. The larger number of thermal complaints in smaller cities could appear counterintuitive considering the presence of urban heat islands in larger cities. However, the result could be driven by better building controls in big cities, where more recent or renovated buildings could be present compared to smaller cities. Another explanation for this distribution of complaints could be linked to the different types of workplaces present in large cities and outside of them (i.e., more office buildings with acoustic issues in large cities and more warehouses with thermal issues in less populated areas). As previously reported, the lack of control of job type in the present analysis is a limitation of this study that should be addressed in future investigations. Due to the nature of the investigated dataset containing a large number of occupants’ feedback across several locations, this is the first time this type of result is presented. Further investigations should be conducted to confirm the reported results and investigate the reasons explaining the distribution of IEQ complaints according to the area population.

![Figure 6](image.png)

Figure 6. The relative percentage of IEQ complaints per location population and IEQ aspects.

4. Concluding remarks

This study investigates the geographical distribution of the sources of IEQ complaints according to climatic zone, country, and population. The IEQ complaints are scraped from online job reviews publicly posted on the Glassdoor website and are categorized according to the four IEQ aspects and their combination. From the descriptive analysis of this unconventional dataset containing 10,568 feedbacks
about the IEQ of several workplaces, it is observed that thermal aspects are the most reported source of complaints independently of the climatic zone and country. Thermal complaints are more reported in smaller cities compared to bigger ones, in which acoustic, visual, and IAQ complaints prevail. This paper shows the great potential to analyze IEQ aspects of several workplaces offered by User-Generated-Content freely available online. The analysis performed in this study offers a picture of the current IEQ discomfort according to the four IEQ aspects and their combination in different locations and is based on quantitative data only (i.e., the number of IEQ word occurrences in online reviews). A qualitative analysis of the posted reviews would provide further insights into the IEQ complaints according to climatic zone, country, and population. The inclusion of the job type would also be beneficial for better understanding the geographical distribution of IEQ complaints. In addition, the dataset used in this analysis is not fully complete in terms of geographical locations (some of the locations are missing) and only reports English reviews, representing a limitation of this study. However, more extensive exploitation of such data by scientists, practitioners, and policymakers could further improve the quantity and quality of reviews posted online. The results reported in this paper and the implementation of the proposed method to analyze different datasets (e.g., referring to smaller organizations or reporting reviews in other languages) can be used to detect the IEQ issues of specific geographic locations to improve the working conditions of several workplaces across the world through more targeted research investigations, policies, and designs.

References

[1] Li P, Froese TM, Brager G. Post-occupancy evaluation: State-of-the-art analysis and state-of-the-practice review. Build Environ 2018;133:187–202. https://doi.org/10.1016/j.buildenv.2018.02.024.

[2] Jayathissa P, Quintana M, Abdelrahman M, Miller C. Humans-as-a-Sensor for Buildings—Intensive Longitudinal Indoor Comfort Models. Buildings 2020;10:174. https://doi.org/10.3390/buildings10100174.

[3] Sheikh Khan D, Kolarik J, Weitzmann P. Design and application of occupant voting systems for collecting occupant feedback on indoor environmental quality of buildings – A review. Build Environ 2020;183:107192. https://doi.org/10.1016/j.buildenv.2020.107192.

[4] Qi M, Li X, Zhu E, Shi Y. Evaluation of perceived indoor environmental quality of five-star hotels in China: An application of online review analysis. Build Environ 2017;111:1–9. https://doi.org/10.1016/j.buildenv.2016.09.027.

[5] Villeneuve H, O’Brien W. Listen to the guests: Text-mining Airbnb reviews to explore indoor environmental quality. Build Environ 2020;169:106555. https://doi.org/10.1016/j.buildenv.2019.106555.

[6] Allen JG, Macomber JD. Healthy buildings: How indoor spaces drive performance and productivity. Harvard University Press; 2020.

[7] Chinazzo G. Investigating the Indoor Environmental Quality of Different Workplaces through Web-scraping and Text-mining of Glassdoor Reviews In preparation.

[8] Glassdoor. About Us. Glas Us n.d. https://www.glassdoor.com/about-us/ (accessed August 14, 2020).

[9] Kahle D, Wickham H. ggmap: Spatial Visualization with ggplot2. R J 2013;5:144–61.

[10] Rubel F, Brugger K, Haslinger K, Auer I. The climate of the European Alps: Shift of very high resolution Köppen-Geiger climate zones 1800–2100. Meteorol Z 2017:115–25. https://doi.org/10.1127/metz/2016/0816.

[11] South A. rworldmap: a new R package for mapping global data. R J 2011;3.

[12] Bryant C, Wheeler NR, Rubel F, French RH. KGC: Koeppen-Geiger climatic zones. R Package Version 1002 2017:1.

[13] Deckmyn MA, Wilks AR. maps: Draw Geographical Maps. R Package Version 330 2018.