


RESEARCH

Open Access



Using data from online geocoding services for the assessment of environmental obesogenic factors: a feasibility study

Maximilian Präger^{1,2}, Christoph Kurz^{1,2}, Julian Böhm^{1,2}, Michael Laxy^{1,2} and Werner Maier^{1,2*} 

Abstract

Background: The increasing prevalence of obesity is a major public health problem in many countries. Built environment factors are known to be associated with obesity, which is an important risk factor for type 2 diabetes. Online geocoding services could be used to identify regions with a high concentration of obesogenic factors. The aim of our study was to examine the feasibility of integrating information from online geocoding services for the assessment of obesogenic environments.

Methods: We identified environmental factors associated with obesity from the literature and translated these factors into variables from the online geocoding services Google Maps and OpenStreetMap (OSM). We tested whether spatial data points can be downloaded from these services and processed and visualized on maps. True- and false-positive values, false-negative values, sensitivities and positive predictive values of the processed data were determined using search engines and in-field inspections within four pilot areas in Bavaria, Germany.

Results: Several environmental factors could be identified from the literature that were either positively or negatively correlated with weight outcomes in previous studies. The diversity of query variables was higher in OSM compared with Google Maps. In each pilot area, query results from Google showed a higher absolute number of true-positive hits and of false-positive hits, but a lower number of false-negative hits during the validation process. The positive predictive value of database hits was higher in OSM and ranged between 81 and 100% compared with a range of 63–89% for Google Maps. In contrast, sensitivities were higher in Google Maps (between 59 and 98%) than in OSM (between 20 and 64%).

Conclusions: It was possible to operationalize obesogenic factors identified from the literature with data and variables available from geocoding services. The validity of Google Maps and OSM was reasonable. The assessment of environmental obesogenic factors via geocoding services could potentially be applied in diabetes surveillance.

Keywords: Obesogenic environment, Geocoding services, Validation, Diabetes

*Correspondence: werner.maier@helmholtz-muenchen.de

¹ Institute of Health Economics and Health Care Management, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, 85758 Neuherberg, Germany
Full list of author information is available at the end of the article



Background

Obesity, commonly defined as a body mass index (BMI) of ≥ 30 kg/m² in adults [1], is the result of a complex multifactorial relationship (e.g. genetic, socioeconomic, and cultural factors) [2]. The prevalence of obesity is affected by lifestyle habits, consumption patterns as well as the urban development [2]. Since the 1980s, the prevalence of obesity has risen considerably and doubled in many countries [3]. Furthermore, a high BMI seems to be associated with a significant proportion of mortality and disability cases [4, 5]. Obesity is therefore recognized as a serious worldwide epidemic.

A number of severe health conditions are correlated with being very overweight, e.g. cardiovascular disease and hypertension, but in particular type 2 diabetes mellitus (T2DM) [6], which is the second leading cause of BMI-related deaths in 2015 [4]. Furthermore, obesity and overweight are the single most relevant predictors for T2DM [7]. Because some studies revealed the simultaneous spread of obesity and diabetes, the term 'diabesity' has been used in the literature in order to illustrate the close connectedness [8].

The built environment, comprising buildings, spaces and products generated or influenced by humans, has a strong influence on promoting or preventing diseases [9, 10]. The built environment can act on three different scales: the macro level describes the sprawl or the compactness of a region on a higher aggregated level, e.g. at the nationwide level, whereas the meso level is concerned with the community or neighbourhood environment, in which the access to certain facilities is of major interest. The micro level constitutes a person-related perspective, for example regarding qualities of urban design, and is often connected with the concept of walkability [11]. Factors of the built environment may contribute to obesity, for example via the availability of unhealthy food or the absence of green spaces [12], and consequently create obesogenic environments. Following Swinburn and colleagues [13], obesogenic environments can be described as 'the sum of influences that the surroundings, opportunities, or conditions of life have on promoting obesity in individuals or populations'.

In order to evaluate features of the built environment, tools based on the use of geographic information systems (GIS) have been developed using remote sensing techniques applicable as desk-based approaches [14]. In the past, researchers have shown great interest in commercial data within GIS-based analyses [15, 16]. Recently, freely available data from online geocoding services such as Google Maps and OpenStreetMap (OSM) have become increasingly popular [17, 18]. These services are often accessed via embedded application programming interfaces (APIs) to search data within the geographical

databases, e.g. for food-related data [19]. These freely available data can be further applied to assess the environmental risk of the development of obesity by describing high- and low-risk geographical areas originating from the accumulation of obesogenic and protective environmental factors [20]. Further applications of such data could refer to environmental pollution or geographical access to primary health care [21, 22].

The aim of our study was to examine the feasibility of integrating information from online geocoding services into the assessment of environmental obesogenic factors which could potentially be used for diabetes surveillance. Diabetes risk has often been estimated e.g. using data from national surveys, but also from administrative data [23]. Thus secondary data from online geocoding services could be a potential complementary data source for diabetes surveillance. Considering this, two steps were required: First, we checked whether obesogenic and protective factors can be derived from the literature and translated into variables from online geocoding services. Second, we compared Google Maps and OSM regarding their validity and reliability of queried data.

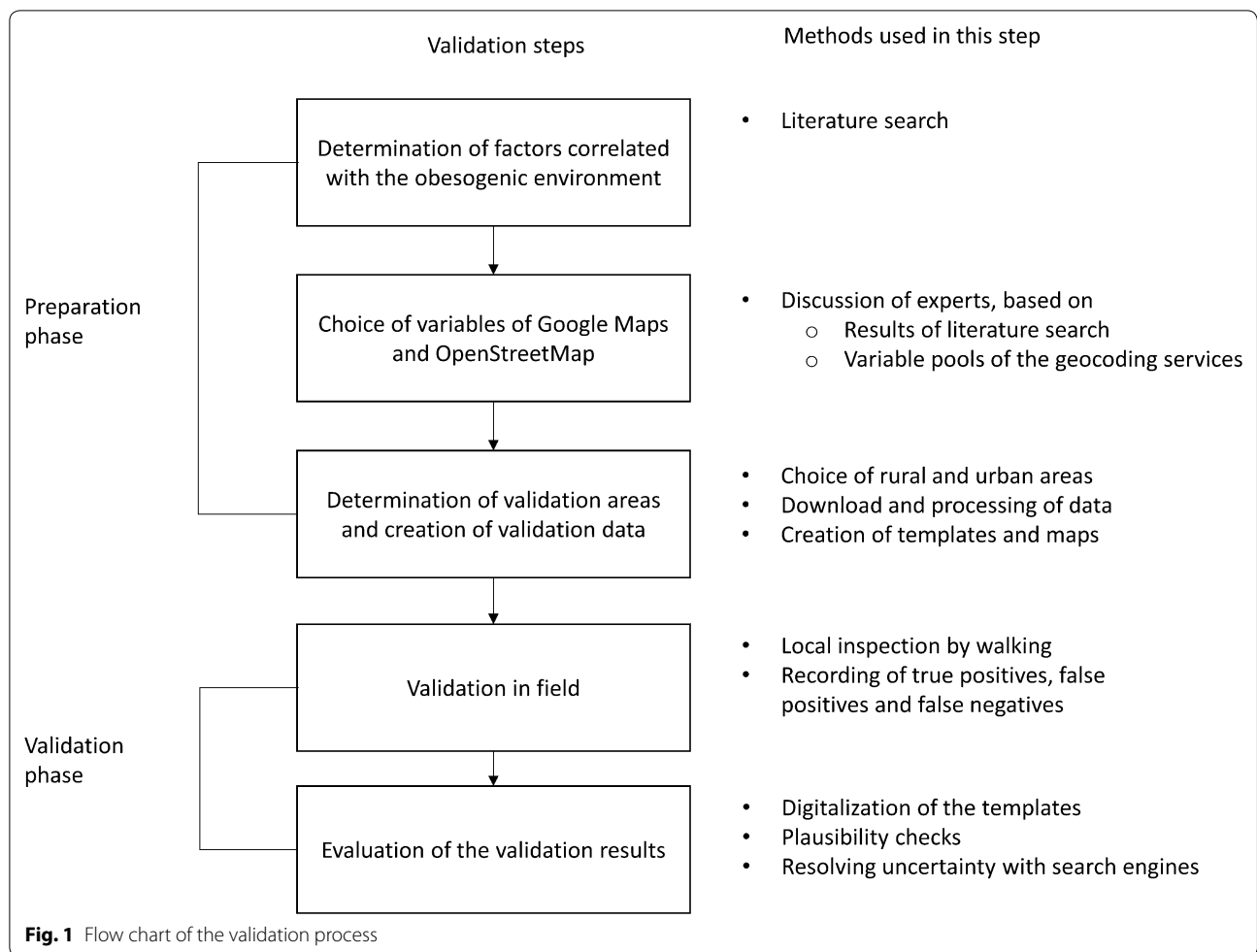
Methods

Design of the validation process

To prepare subsequent validations, we initially identified environmental factors correlated with obesity from the literature. Based on these results and on expert discussions, we have chosen variables from Google Maps and OSM and downloaded these for four regions in Bavaria, Germany. Subsequently, these downloaded data points were validated in the field and by using search engines. An overview of the methods applied during the two phases of preparation and validation is shown in Fig. 1, and further details are provided below.

Literature search and extraction of variables

We applied a search strategy within PubMed using the search terms 'obesogenic', 'environmental factors', 'systematic' and 'review'. After screening the results, two reviews were determined to be relevant for our analysis. The first review by Mackenbach and colleagues [24] provided a systematic search strategy and identified correlates of environmental factors with obesity. The second publication was a review of GIS methods by Jia and colleagues [25], in which correlations of variables with weight status and obesity were described. Following Mackenbach et al. [24], we created a table in order to summarize the factors from our literature search. In a first step, we extracted environmental factors from the studies covered by the two reviews. In a second step, we grouped the publications describing these environmental factors and extracted and summarized information from



these publications in order to determine their correlation with obesity.

Subsequently, we extended the systematic search strategy provided by Mackenbach et al. in order to identify recent additional studies within PubMed, EMBASE, Web of Science, Cochrane Library, PsychInfo and Google Scholar. We completed the variable table with the additionally identified publications, and information from these studies was used to update the correlations of the environmental factors.

Definition of correlation

For each given environmental factor, we summed up the numbers of studies describing a positive and significant correlation with obesity. Analogously, we counted the numbers of studies describing negative and significant correlations with obesity for the same given factor. Subsequently, we defined this factor as overall positively correlated if at least three publications could be found and if the ratio of the number of positive correlations for the factor divided by the number of negative correlations

for the same factor was 2 or higher. Dividing by 0 in this sense can be interpreted as causing infinity. Studies showing no significant correlation were not taken into account. Analogously, if the number of negative correlations divided by the number of positive correlations equals 2 or more, we assumed the factor to be overall negatively correlated. Otherwise, we supposed that no association existed. For example, if a factor was described with a positive correlation in five publications and with a negative correlation in 12 publications, an overall negative correlation was assumed as $12/5 \geq 2$. This calculation procedure was performed for each extracted environmental factor.

Determining the variables from the geocoding services

We checked environmental factors identified within the literature search regarding mapping possibilities with variables from Google Maps and OSM. Google Maps data, among other sources, are derived from official registries, e.g. from the Agency for Digitisation, High-Speed Internet and Surveying in Bavaria [26, 27]. OpenStreetMap,

in contrast, is based on volunteered geographical information (VGI), i.e. it is based on user-generated content [28]. We have chosen both geocoding services because their data were freely available at low cost. Furthermore, their accuracy has been investigated for Germany in the past. Apparently, Google Maps showed higher completeness and higher precision of coordinates than OSM [17]. Besides environmental obesogenic factors, additional variables concerning the regional healthcare structure were taken into account. Four researchers in our team independently rated the relevance of the variables from the geocoding services with respect to the results of the literature search. After discussion, the variables best operationalizing the identified factors from the literature were determined and downloaded from Google Maps and OSM. We focused our analysis on single points of interest (POIs). Therefore, complex variable constructs, such as 'neighbourhood walkability' and 'land use mix', were not considered, as these compound measures are based e.g. on residential density or numbers of developed hectares which cannot be directly derived from online geocoding services. For an overview on the composites of these variables see Feng et al. [29]. Furthermore, six broader categories, 'food', 'doctor', 'sport', 'education', 'transport' and 'other', were determined via expert discussions within our team, and each operationalized variable, for which POIs were returned by at least one geocoding service, was assigned to one of those categories. Based on this approach, it was possible to visualize the distribution of environmental factors on a higher aggregated level and improve interpretability of the field validation results.

Choosing locations for the validation process

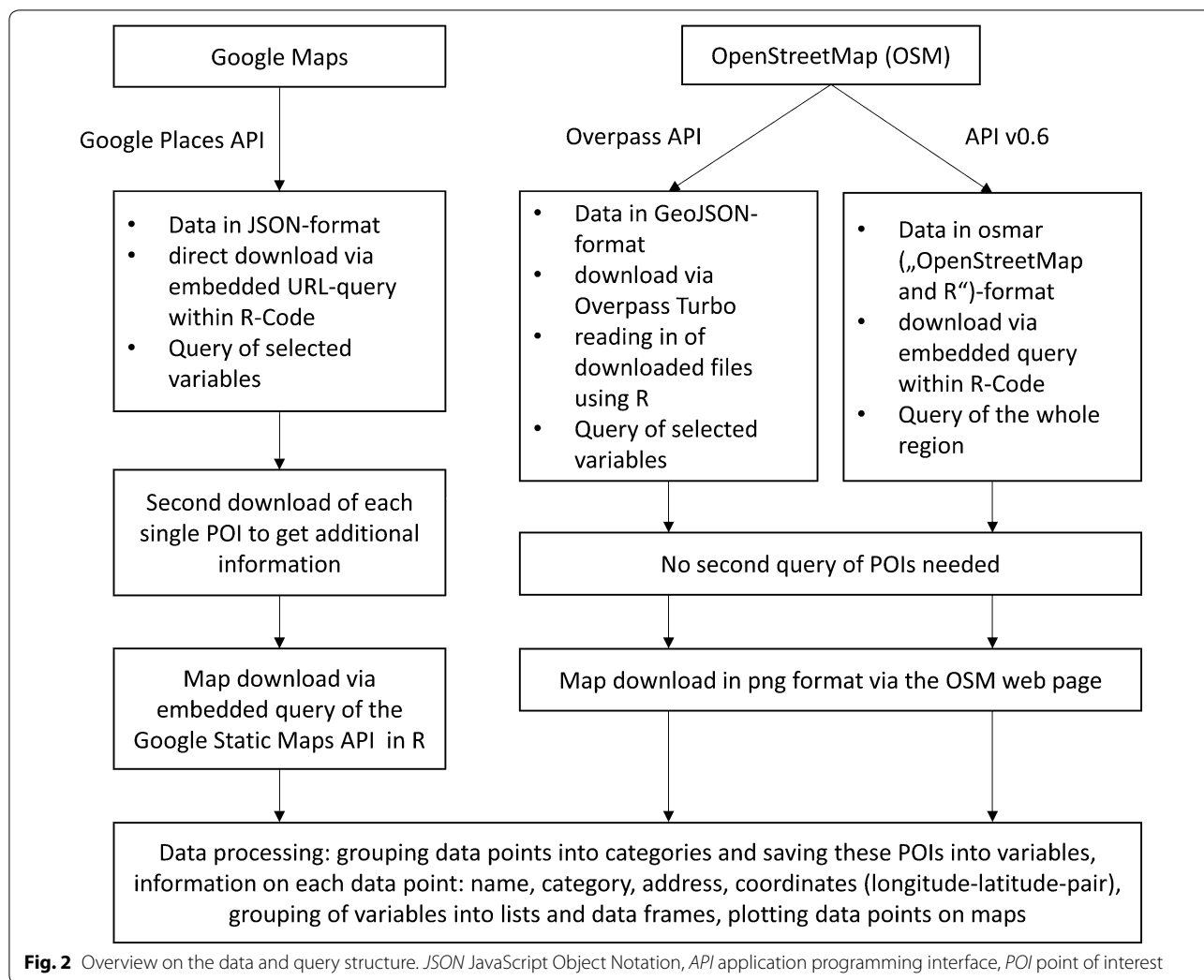
We have chosen four pilot areas in the German federal state of Bavaria for the validation process. Our aim was to investigate the data quality of the geocoding services within regions of different population density and urbanization level. Size and population count of each area were derived from the German Federal Statistical Office and the statistical offices of the German Länder (federal states) [30]. The first region was a sparsely populated municipality in the south-west of Bavaria with fewer than 2000 inhabitants encompassing an area of about 36.31 km². This area constituted a rural region containing few amenities (Area A). The second area was a street in a medium-sized major district town near Munich, the capital of Bavaria, with fewer than 45,000 inhabitants and a size of around 34.96 km² (Area B). Finally, the densely populated city of Munich (about 1.5 million inhabitants, total area 310.71 km²) was selected for the validation. From the whole city of Munich, both a denser area close to the city centre and an area with a relatively lower density of amenities was chosen (Areas C and D).

Database extraction and processing of data

Google Maps data were downloaded using queries in uniform resource locator (URL) format targeting the Google Places API. Furthermore, OSM queries were performed using a web interface and an OSM-based R package. The geographical database returns data in JavaScript Object Notation (JSON), GeoJSON or osmar ('OpenStreetMap and R') format, which are standard representations for geographical data. Based on the structure of these data formats, information for each of the single POIs can be accessed efficiently via a hierarchical structure and subsequently processed. For the Google Maps results, each entry had to be queried again in order to get additional relevant information, e.g. on names, addresses and categorizations. Using the downloaded OSM data, additional information could be extracted directly from the previously described data formats without any additional query. An overview of the data formats and query possibilities is shown in Fig. 2. The return of the spatial databases was checked regarding consistency and plausibility. Important examinations were identifying POIs that were counted twice or more because of being listed within different categories and checking whether the return of the database lies completely within the pre-specified search area. Additionally, spatial POIs were visualized on maps in order to check coherence. The geographical data points were marked according to their factor category, and the search area was also plotted. An example of visualization of some factors for Area D can be found in Fig. 3 for OSM. The underlying code and the other codes regarding Area D are available on github [<https://github.com/MAPraeger/GOcode>. Accessed 23 April 2019].

Search area and download capacity

The shape of the downloaded regions was predefined by the geocoding services. OSM areas were rectangular, whereas Google Maps areas were circular. In order to make the shapes of OSM and Google Maps queries more comparable, we defined OSM search regions as quadratic. Further differences between the geocoding services affected the maximum downloadable data size. At the time of data download in 2017, Google Maps allowed up to 200 results per query and 1000 queries per day per person at zero costs [31], whereas OSM had fewer restrictions [32]. Depending on the API and the download tools used, areas of arbitrary size, whole so-called 'planet files' [33] or nearly arbitrary data sizes caused by memory overload within the statistical software, could be downloaded. Therefore, areas for the validation process were determined such that none of the above-mentioned restrictions took effect. Owing to the lower number of spatial POIs within the rural region (Area A), a wider area containing the whole municipality was chosen compared



with the more urban areas (Areas B–D), for which the diameters of the circles and the edges of the squares were set to 200 metres.

Validation process

Four researchers in our team locally scanned the pre-defined validation areas looking for the existence of the downloaded POIs of Google Maps and OSM. We designed a template to standardize the recording process and used maps containing the data points to improve efficiency. The number of returned POIs of a database was called ‘hits’. Each researcher documented the validation date, confirmation (true positive hit) or rejection (false positive hit) of existence of the POIs and new record of false negatives, i.e. data points discovered in the field that were not covered by Google Maps or OSM or both. After completion, the templates were digitalized.

If uncertainties regarding the existence of a POI were present during validation in the field, the researchers

recorded their comments. If these notes indicated restrictions, e.g. regarding access to certain facilities during in-field validations, several online search engines were used to resolve these uncertainties. Further examples were incorrect categorization or implausible numbers of false positives at a certain place. To overcome these issues, we visited the home pages of the affected amenities and considered business directories (yellow pages).

Common summary statistics for the validation of geographical data points were calculated. For the quality assessment of the performance of a geocoding service for a given area, sensitivities, i.e. true positives divided by the sum of true positives and false negatives, and positive predictive values (PPVs), i.e. true positives divided by the sum of true positives and false positives, were calculated [34, 35].

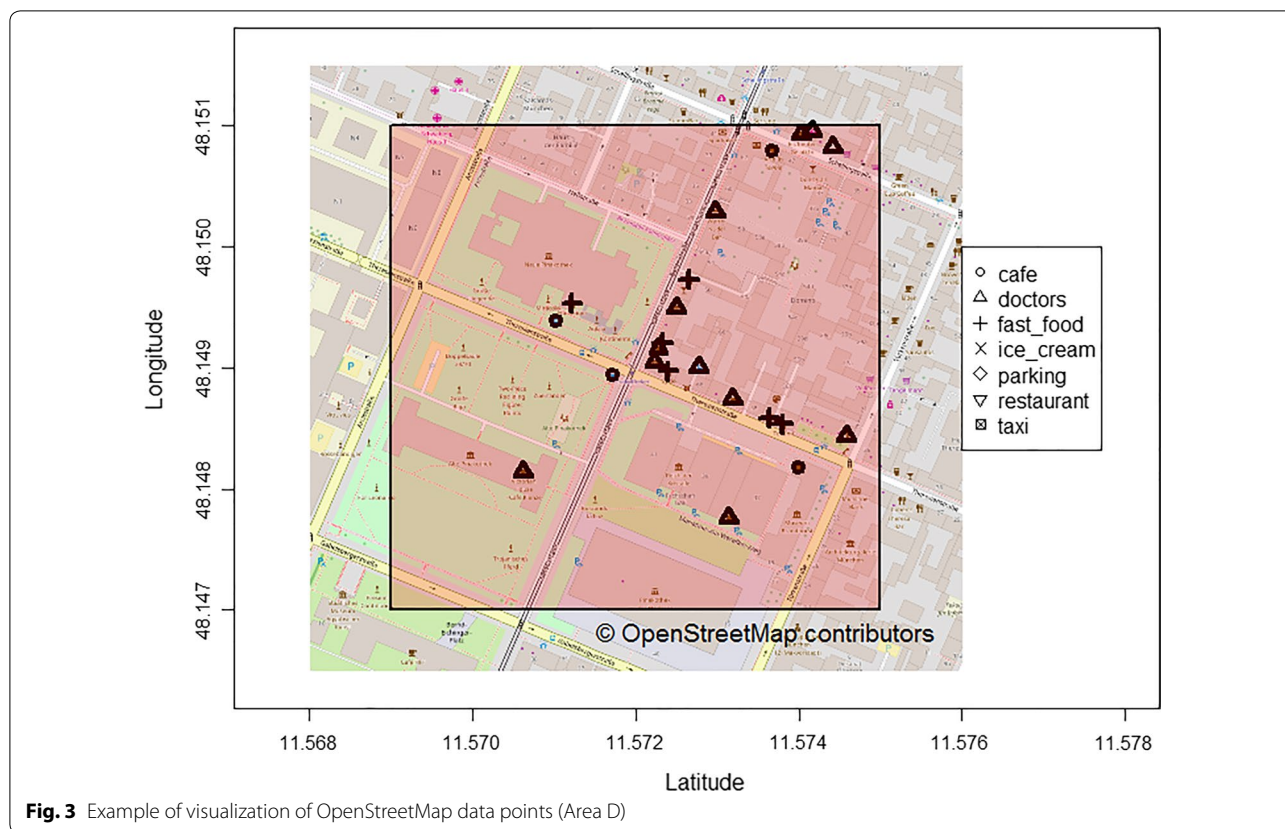


Fig. 3 Example of visualization of OpenStreetMap data points (Area D)

Software

We used the free software environment R, version 3.3.2, to implement code targeting the Google Places API via embedded URL query and for processing of the query results [36]. In order to download data from OSM, we applied an online tool for data filtering (Overpass Turbo) and the R package ‘osmar’ [37, 38]. For data processing, we used the packages ‘geojsonR’, ‘jsonlite’ and ‘rgdal’ and, for data visualization, the R packages ‘ggmap’ and ‘ggplot2’ [39–43].

Results

Literature search

An extensive list of environmental factors and the corresponding references (N = 256) can be found within Additional file 1: Table S1. The table contains the numbers of studies describing positive correlations, negative correlations and studies without significant associations for a given environmental factor. According to the definition of correlation within the methods section, overall positive correlations with weight status were discovered for the variables ‘fast food’, ‘food retail’, ‘unhealthy food outlets’, ‘convenience store’, ‘rural areas’, ‘urban sprawl’, ‘county sprawl’, ‘traffic’, ‘transport’ and ‘poverty’. Overall negative

correlations were found for the variables ‘(healthy) food outlets’, ‘restaurants’, ‘supermarkets’, ‘tree cover’, ‘fitness or physical activity facilities’, ‘forests’, ‘greenspace’, ‘longer way to school’, ‘open space’, ‘outdoor recreation’, ‘park’, ‘recreation centre’, ‘walkability’, ‘aesthetics’, ‘intersection density’, ‘land use mix’, ‘population density’, ‘safety’, ‘side-walk completeness’, ‘street connectivity’, ‘education’ and ‘physician supply’.

Chosen variables from Google Maps and OSM

Tables 1 and 2 show the factors from Google Maps (N=25 in total) and OSM (N=126 in total) chosen for the validation process. Owing to the extent of the OSM variable pool, the relevant factors in the category

Table 1 Selected variables from the Google Maps pool

Bakery	Bar	Bus station	Cafe
Convenience store	Dentist	Doctor	Food
Grocery or supermarket	Gym	Hospital	Meal delivery
Meal takeaway	Park	Pharmacy	Physiotherapist
Restaurant	School	Spa	Stadium
Subway station	Taxi stand	Train station	Transit station
University			

Table 2 Selected OpenStreetMap (OSM) variables in the category ‘amenity’

Bar	Bbq	Biergarten	Cafe	Fast food
Food court	Ice cream	Pub	Restaurant	College
School	Bicycle parking	Bicycle rental	Boat sharing	Bus station
Taxi	Clinic	Dentist	Doctors	Hospital
Nursing home	Pharmacy	Dive centre	Dojo	Ranger station
Beach resort	Dance	Fishing	Fitness centre	Garden
Golf course	Ice rink	Nature reserve	Park	Pitch
Playground	Sports centre	Stadium	Swimming area	Swimming pool
Track	Water park			

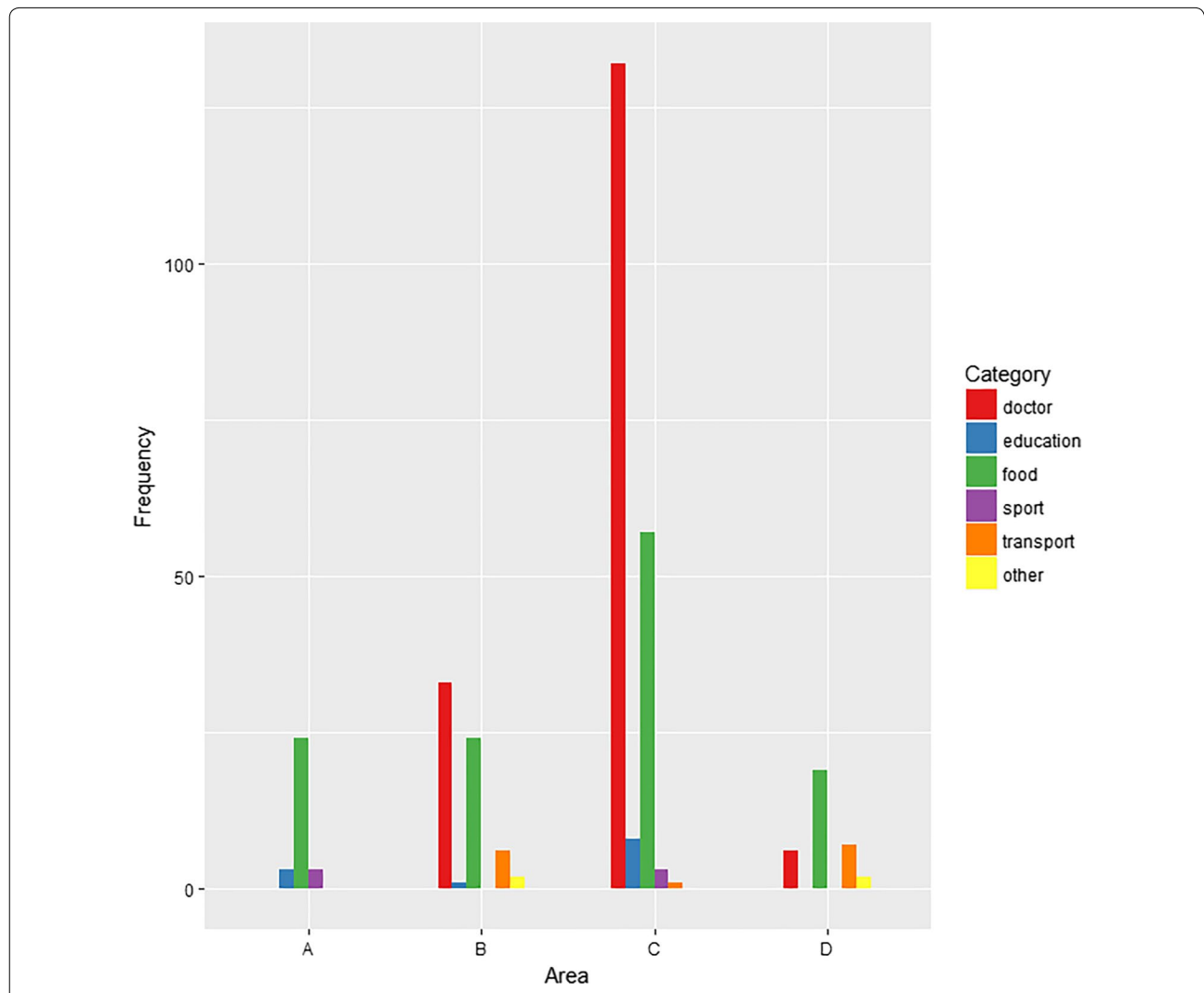


Fig. 4 Distribution of hits across variable categories using Google Maps. Area A: sparsely populated municipality in the south-west of Bavaria. Area B: street in a medium-sized populated major district town near Munich. Area C: area close to the centre within the densely populated city of Munich. Area D: area with a lower density of amenities within the densely populated city of Munich

'amenity' are shown within Table 2 (N=42). The full list of OSM variables is shown in Additional file 2: Table S2.

Distribution of database results across categories

Bar charts are shown for Google Maps (Fig. 4) and OSM (Fig. 5) in order to visualize the distribution of the database hits, i.e. the distribution of the sum of true-positive and false-positive entries of the geocoding services, across the six categories of 'doctor', 'education', 'food', 'sport', 'transport' and 'other' for the validation areas. Within the medium-sized populated Area B and the densely populated Area C, predominantly entries in the categories 'doctor' and 'food' account for most of the database hits in Google Maps. For the remaining areas, using Google Maps, 'food' was the most relevant

category. Regarding OSM, the category 'food' was the most frequent category within Areas C and D of the city of Munich.

Validations

Tables 3 and 4 show the numbers of true positives and false positives, PPVs, numbers of false negatives and sensitivity values for each validation area. As shown in the table, absolute numbers of true hits were higher for Google Maps than the corresponding numbers for OSM, irrespective of the validation area under consideration. Furthermore, false positives were also higher for Google Maps compared with OSM. The PPVs of OSM hits, ranging between 81 and 100%, were higher than the PPVs of Google Maps hits, which were found to be between 65

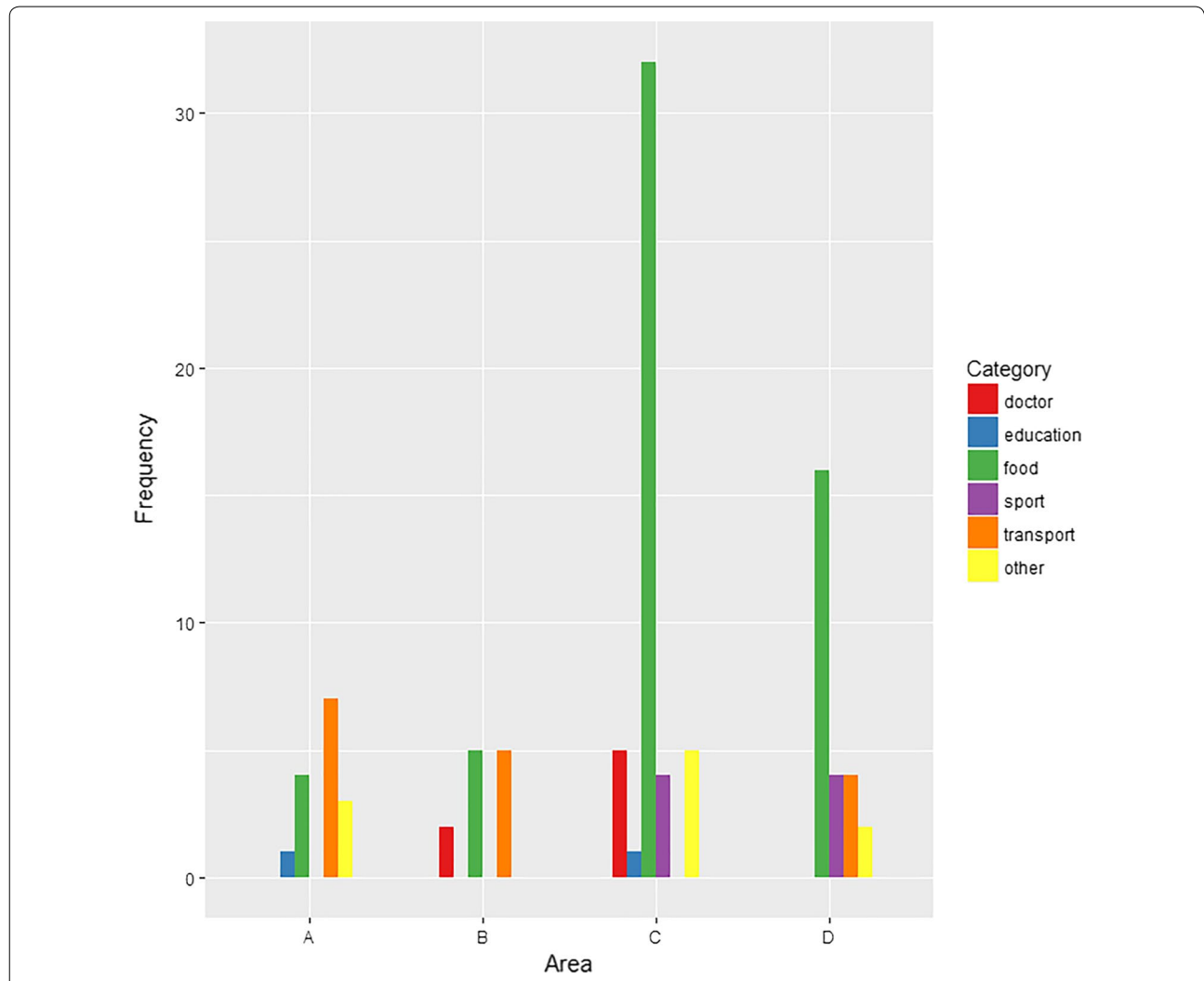


Fig. 5 Distribution of hits across variable categories using OpenStreetMap. Area A: sparsely populated municipality in the south-west of Bavaria. Area B: street in a medium-sized populated major district town near Munich. Area C: area close to the centre within the densely populated city of Munich. Area D: area with a lower density of amenities within the densely populated city of Munich

Table 3 Results of the field validation

Area	Geocoding service	True positives: N (% positive) ^a	False positives: N (% positive)	False negatives: N	Sensitivity ^b : %
A	Google Maps	19 (63.33)	11 (36.67)	13	59.38
A	OpenStreetMap	15 (88.24)	2 (11.76)	17	46.88
B	Google Maps	58 (89.23)	7 (10.77)	1	98.31
B	OpenStreetMap	12 (100)	0 (0)	47	20.34
C	Google Maps	144 (71.64)	57 (28.36)	63	69.57
C	OpenStreetMap	41 (87.23)	6 (12.77)	166	19.81
D	Google Maps	22 (64.71)	12 (35.29)	11	66.67
D	OpenStreetMap	21 (80.77)	5 (19.23)	12	63.64

Area A: sparsely populated municipality in the south-west of Bavaria

Area B: area in a medium-sized populated major district town near Munich

Area C: area close to the centre within the densely populated city of Munich

Area D: area with a lower density of amenities within the densely populated city of Munich

^a The percentage of true positives is the positive predictive value (PPV) [PPV = true positives/(true positives + false positives)]

^b Sensitivity = true positives/(true positives + false negatives)

Table 4 Results of the field validation without the category 'doctor'

Area	Geocoding service	True positives: N (% positive) ^a	False positives: N (% positive)	False negatives: N	Sensitivity ^b : %
A	Google Maps	18 (62.07)	11 (37.93)	13	58.06
A	OpenStreetMap	15 (88.24)	2 (11.76)	16	48.39
B	Google Maps	29 (90.63)	3 (9.38)	1	96.67
B	OpenStreetMap	10 (100)	0 (0)	20	33.33
C	Google Maps	48 (69.57)	21 (30.43)	30	61.54
C	OpenStreetMap	36 (85.71)	6 (14.29)	42	46.15
D	Google Maps	19 (67.86)	9 (32.14)	6	76.00
D	OpenStreetMap	21 (80.77)	5 (19.23)	4	84.00

Area A: sparsely populated municipality in the south-west of Bavaria

Area B: area in a medium-sized populated major district town near Munich

Area C: area close to the centre within the densely populated city of Munich

Area D: area with a lower density of amenities within the densely populated city of Munich

^a The percentage of true positives is the positive predictive value (PPV) [PPV = true positives/(true positives + false positives)]

^b Sensitivity = true positives/(true positives + false negatives)

and 89%. In contrast, sensitivities were higher in Google Maps (between 59 and 98%) than in OSM (between 20 and 64%). False negatives were higher for OSM within three of the four validation areas. An overall comparison between the four areas showed that Area C within the city of Munich had the highest numbers of false negatives for both geocoding services. For OSM, high numbers of false negatives were also discovered for Area B, i.e. for the major district town. Predominantly during the validation within Area C, it became evident that the data quality regarding the variable category 'doctor' had a fundamental influence on the validation results. Therefore, we recalculated Table 3 without the POIs belonging to this

category. The results of this recalculation process can be found within Table 4. Having omitted the category 'doctor', sensitivities of OSM improved for Area B and Area C. Within Area D, sensitivities of OSM were higher than sensitivities of Google Maps.

Discussion

The aim of our study was to examine the feasibility of integrating information from online geocoding services for the assessment of environmental obesogenic factors that could potentially be used for diabetes surveillance. First, we identified variables correlated with obesogenic environments from the literature. Subsequently, we

tested whether these variables could be reproduced using data from the online geocoding services Google Maps and OpenStreetMap (OSM). The results showed that this was possible given some restrictions, predominantly the diversity of the variable pools of the geocoding services and the complexity of the environmental factor to be projected. Maps created from the obesogenic and from protective data showed the geographical distribution of the environmental factors and were used within subsequent field validations. On the one hand, Google Maps showed greater completeness, i.e. lower proportion of false negatives, regarding POIs subsequently discovered in the field and the additional information assigned to them. Furthermore, the sensitivity of Google Maps was higher than the sensitivity of OSM. On the other hand, a higher PPV was seen for OSM in each of the validation areas.

Recently, the validity of the geocoding service Google Maps was tested using geoprocessing information [18]. Instead of using single geographical data points from the spatial databases of Google Maps and OSM, the authors compared virtual audit via Google Street View. Additionally, local field inspections were performed as the gold standard. It was shown that the validity and reliability of using Google Maps for the assessment of the built environment was high (Kappa of 78% and 80% respectively). Considering the German context, field inspections concerning the obesogenic environment have been performed in the past in order to record POIs [44]. Therefore, it was an important step within our study to inspect the database results of Google Maps and OSM locally.

PPVs of Google Maps and OSM found during our validation process were compared with each other. It became evident that the PPV of OSM was higher than the PPV for Google Maps in each region, because Google Maps showed considerably more false positives. Considering sensitivity, OSM showed lower values than Google Maps. Most influential variables regarding these comparisons were found within the category 'doctor'. The data quality regarding physicians was better for Google Maps compared to OSM. Therefore, within areas with a higher share of doctors (Area B and Area C) the differences in sensitivities between Google Maps and OSM were large. Deleting the category 'doctor' from the analysis thus moderated this difference. False positives of Google Maps within the densely populated Area C were also mainly caused by the category 'doctor'. The same category also contributed to the number of false negatives in OSM within this area and the sensitivity of OSM improved considerably after omitting POIs belonging to this category (see Table 4). To highlight the different influences of certain variable groups, it was an important step in our validation process to look for suitable stratification

structures, such as the six categories 'doctor', 'food', 'sport', 'transport', 'education' and 'other'.

In our study, we calculated the sensitivities and PPVs of Google Maps and OSM hits. They can be compared with the PPVs of other POI databases that we found in the existing literature. For example, Clary and colleagues [34] validated a Canadian food outlet database in the field. Comparing their database results with the actual occurrences in the field, the authors found sensitivities between 54.5 and 65.5% as well as PPVs between 64.4 and 77.3%. Within our study, the PPV for OSM was markedly higher (between 81 and 100%), whereas Google Maps had a more similar PPV compared with the Canadian database (between 63 and 89%). Regarding sensitivity, OSM showed lower values within three of the four validation areas (between 20 and 64%), whereas Google Maps sensitivities were at least comparable (between 59 and 98%) with the food outlet database.

To evaluate features of the obesogenic environment, Bethlehem and colleagues [14] performed a virtual audit based on Google Earth (GE) and Google Street View (GSV). They assessed the aspects walking, cycling, public transport, aesthetics, land use mix, grocery stores, food outlets and recreational facilities using observers. Virtual audit was found to be a valid and reliable approach. Within our study, we used Google Maps and OSM APIs for the programmed download of POIs, which does not need individual assessment for data collection.

Within our analyses, it also became evident that new variable entries appear more frequently, but old entries were deleted with time lag within the Google Maps database. The more specific variables in the OSM pool made it possible to identify some POIs that could not be precisely queried by Google. For example, OSM made it possible to extract 'fast food' instead of the broader category 'food'. This feature nevertheless required taking into account all relevant specific factors describing a variable at a higher level in order to exhaust the OSM database completely.

Strengths and limitations

Our study is based on an extensive literature search extracting factors of obesogenic environments. We used freely available data from global geocoding services Google Maps and OSM and applied various methods for downloading and processing geographical data using new query codes in the R programming environment. Finally, we validated our results with in-field inspections. To evaluate both physical activity and food-related environmental factors, composite approaches are required, which have been performed rather infrequently in the past [12]. Within our approach, we combined the food environment and the physical activity environment into

a single layer containing POIs of the obesogenic factors and POIs of the protective factors.

Some limitations of our study have to be mentioned. First, it is focused on evidence from the literature based on an energy imbalance model [45]. However, according to the recent literature, other etiological causes for the development of obesity have to be considered to fully understand the underlying mechanisms, e.g. the carbohydrate-insulin model of obesity (beyond ‘calories in, calories out’) [46] or dietary behaviour (‘ultra-processed food vs unprocessed food’) [47]. Second, the literature search had a broad scope by updating and complementing a systematic review; however, a large number of the identified studies originated from the US. Structural differences regarding the built environment in US and European cities may influence direct transferability to the European context. For example, cities in the US are much more car dependent than European cities, which results in expected different health effects of environmental factors associated with physical activity [48]. Furthermore, instead of unhealthy corner stores in the US, in European countries, healthy stores selling fresh fruit and vegetables exist more often and are more evenly distributed across the cities [49]. A third drawback regarding our literature search could be publication bias, which would influence the assessment of the overall correlation of an environmental factor [50, 51]. Fourth, a significant proportion of the environmental factors was not correlated with obesity in the same direction across studies. Given this restriction, we have summarized the correlations found in the literature based on expert decision. Fifth, the precision and feasibility of variable extraction fundamentally depend on the variable pool structure of the geocoding service. Differences in the definition of a variable across geocoding services hamper direct comparisons of variables. Within our study, we found that the variable pool of OSM contains many more variables than Google Maps for a large number of environmental factors. Finding broader categories for environmental factors within our analysis made it easier to compare variables across geocoding services. Sixth, our study was limited to a German environment; therefore, generalization of our findings needs further assessment in other countries. Seventh, we have downloaded spatial POIs at a certain point in time; thus, we cannot make inferences on time effects. However, this cross-section offers an important starting point for future analyses. Eighth, each geographical area was validated by a different researcher; therefore, interobserver variability could have appeared during validation. In order to counteract this kind of bias, prior instructions were defined as precisely as possible, and discussions between the observers took place both before and after the validations. Finally, some restrictions regarding

access to certain facilities appeared during validations in the field, mostly concerning database hits of the category ‘doctor’. Results of the validation process without this category are shown in Table 4. Within the analysis including the category ‘doctor’, this generated some uncertainties; therefore, we used the best available evidence, i.e. the home pages of these amenities and business directories (yellow pages). However, these uncertainties occurred only in a small number of cases and were discussed in detail during processing of the validation results.

The aim of our study was to examine the feasibility of using data from online geocoding services for diabetes surveillance. We were able to integrate information from these services by downloading, processing and visualizing their data on maps. The reliability of these variables was assessed within field validations and by search engines. Future examinations could test further types of variables from other research areas.

Conclusions

Based on an extensive literature search, environmental factors could be identified that are associated with obesity. These factors could be partly operationalized through the variables and data available from the online geocoding services Google Maps and OSM. Using APIs, spatial data points could be identified and subsequently visualized on maps. Our findings showed that the validity of data from online geocoding services was reasonable. Consequently, environmental obesogenic factors could be described with our methodology and potentially used within diabetes surveillance. Further validation studies are needed to investigate the importance of environmental obesogenic factors.

Additional files

Additional file 1: Table S1. Factors determined by literature search.

Additional file 2: Table S2. Complete list of chosen OSM variables.

Abbreviations

API: application programming interface; BMI: body mass index; GE: Google Earth; GIS: geographic information system; GSV: Google Street View; JSON: JavaScript Object Notation; OSM: OpenStreetMap; POI: point of interest; PPV: positive predictive value; T2DM: type 2 diabetes mellitus; URL: uniform resource locator; VGI: volunteered geographical information.

Acknowledgements

Not applicable.

Authors' contributions

WM initiated the cooperation and supervised the project. MP and JB were responsible for the literature search and extraction of environmental factors. MP programmed the database queries, processed the data and visualized the results on maps. CK provided statistical input. ML and WM provided health scientific input. MP, JB, CK and ML performed field validations. MP wrote the manuscript. All authors read previous versions of the manuscript, commented

on and approved the final version of the manuscript for submission. All authors read and approved the final manuscript.

Funding

The study was supported by the Federal Ministry of Health, Germany (FKZ: GE20160324) as part of the German Diabetes Surveillance Project conducted by the Robert Koch Institute, Berlin.

Availability of data and materials

The data generated for this study were downloaded from Google Maps and OSM and the exemplary code is available on github [<https://github.com/MAPraeger/GOcode>. Accessed 23 April 2019] as described in the methods section.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Institute of Health Economics and Health Care Management, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, 85758 Neuherberg, Germany. ² German Center for Diabetes Research, Neuherberg, Germany.

Received: 22 January 2019 Accepted: 29 May 2019

Published online: 07 June 2019

References

- Ogden CL, Yanovski SZ, Carroll MD, Flegal KM. The epidemiology of obesity. *Gastroenterology*. 2007;132(6):2087–102.
- Apovian CM. Obesity: definition, comorbidities, causes, and burden. *Am J Managed Care*. 2016;22(7 Suppl):s176–85.
- Gregg EW, Shaw JE. Global health effects of overweight and obesity. *N Engl J Med*. 2017;377(1):80–1.
- Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, Lee A, Marczak L, Mokdad AH, Moradi-Lakeh M, Naghavi M, et al. Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med*. 2017;377(1):13–27.
- Alam S, Lang JJ, Drucker AM, et al. Assessment of the burden of diseases and injuries attributable to risk factors in Canada from 1990 to 2016: an analysis of the Global Burden of Disease Study. *CMAJ Open*. 2019;7(1):E140–8.
- Bischoff SC, Boirie Y, Cederholm T, Chourdakis M, Cuerda C, Delzenne NM, Deutz NE, Fouque D, Genton L, Gil C, et al. Towards a multidisciplinary approach to understand and manage obesity and related diseases. *Clin Nutr*. 2017;36(4):917–38.
- Chen L, Magliano DJ, Zimmet PZ. The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives. *Nat Rev Endocrinol*. 2011;8(4):228–36.
- Verma S, Hussain ME. Obesity and diabetes: an update. *Diabetes Metab Syndr*. 2017;11(1):73–9.
- Bhatnagar A. Environmental determinants of cardiovascular disease. *Circ Res*. 2017;121(2):162–80.
- Brisbon N, Plumb J, Brawer R, Paxman D. The asthma and obesity epidemics: the role played by the built environment—a public health perspective. *J Allergy Clin Immunol*. 2005;115(5):1024–8.
- Garfinkel-Castro A, Kim K, Hamidi S, Ewing R. Obesity and the built environment at different urban scales: examining the literature. *Nutr Rev*. 2017;75(suppl 1):51–61.
- Townshend T, Lake A. Obesogenic environments: current evidence of the built and food environments. *Perspect Public Health*. 2017;137(1):38–44.
- Swinburn B, Egger G, Raza F. Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Prev Med*. 1999;29(6 Pt 1):563–70.
- Bethlehem JR, Mackenbach JD, Ben-Rebah M, Compernelle S, Glonti K, Bardos H, Rutter HR, Charreire H, Oppert JM, Brug J, et al. The SPOTLIGHT virtual audit tool: a valid and reliable tool to assess obesogenic characteristics of the built environment. *Int J Health Geogr*. 2014;13:52.
- Lebel A, Daepf MI, Block JP, Walker R, Lalonde B, Kestens Y, Subramanian SV. Quantifying the foodscape: a systematic review and meta-analysis of the validity of commercially available business data. *PLoS ONE*. 2017;12(3):e0174417.
- Thornton LE, Pearce JR, Kavanagh AM. Using geographic information systems (GIS) to assess the role of the built environment in influencing obesity: a glossary. *Int J Behav Nutr Phys Act*. 2011;8:71.
- Lemke D, Mattauch V, Heidinger O, Hense HW. [Who hits the mark? A comparative study of the free geocoding services of Google and OpenStreetMap]. *Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany))*. 2015;77(8–9):e160–5.
- Silva V, Grande AJ, Rech CR, Peccin MS. Geoprocessing via Google Maps for assessing obesogenic built environments related to physical activity and chronic noncommunicable diseases: validity and reliability. *J Healthc Eng*. 2015;6(1):41–54.
- Seto E, Hua J, Wu L, Bestick A, Shia V, Eom S, Han J, Wang M, Li Y. The Kunming CalFit study: modeling dietary behavioral patterns using smartphone data. In: Conference proceedings: 2014 annual international conference of the IEEE engineering in medicine and biology society. 2014. p. 6884–7.
- Feuillet T, Charreire H, Roda C, Ben Rebah M, Mackenbach JD, Compernelle S, Glonti K, Bardos H, Rutter H, De Bourdeaudhuij I, et al. Neighbourhood typology based on virtual audit of environmental obesogenic characteristics. *Obes Rev*. 2016;17(Suppl 1):19–30.
- Li B, Wang J, Wu S, Jia Z, Li Y, Wang T, Zhou S. New method for improving spatial allocation accuracy of industrial energy consumption and implications for polycyclic aromatic hydrocarbon emissions in China. *Environ Sci Technol*. 2019;53(8):4326–34.
- Schuurman N, Berube M, Crooks VA. Measuring potential spatial access to primary health care physicians using a modified gravity model. *Can Geogr*. 2010;54(1):29–45.
- Ali MK, Siegel KR, Laxy M, Gregg EW. Advancing measurement of diabetes at the population level. *Curr Diabetes Rep*. 2018;18(11):108.
- Mackenbach JD, Rutter H, Compernelle S, Glonti K, Oppert JM, Charreire H, De Bourdeaudhuij I, Brug J, Nijpels G, Lakerveld J. Obesogenic environments: a systematic review of the association between the physical environment and adult weight status, the SPOTLIGHT project. *BMC Public Health*. 2014;14:233.
- Jia P, Cheng X, Xue H, Wang Y. Applications of geographic information systems (GIS) data and methods in obesity-related research. *Obes Rev*. 2017;18(4):400–11.
- Google. Legal notices for Google Maps/Google Earth and Google Maps/Google Earth APIs. https://www.google.com/intl/en_ALL/help/legalnotices_maps.html. Accessed 11 Sept 2018.
- Landesamt für Digitalisierung Breitband und Vermessung. <https://www.lbv.bayern.de/index.html>. Accessed 11 Sept 2018.
- Neis P, Zielstra D. Recent developments and future trends in volunteered geographic information research: the case of OpenStreetMap. *Future Internet*. 2014;6(1):76–106.
- Feng J, Glass TA, Curriero FC, Stewart WF, Schwartz BS. The built environment and obesity: a systematic review of the epidemiologic evidence. *Health Place*. 2010;16(2):175–90.
- Statistische Ämter des Bundes und der Länder. Gemeinsames Statistikportal. Gemeindeverzeichnis-Online. <https://www.statistikportal.de/de/produkte/gemeindeverzeichnis>. Accessed 11 Sept 2018.
- Places API. <https://developers.google.com/places/web-service/intro>. Accessed 29 Aug 2017.
- Downloading data. https://wiki.openstreetmap.org/wiki/Downloading_data. Accessed 27 Nov 2018.
- Planet.osm. <https://wiki.openstreetmap.org/wiki/Planet.osm>. Accessed 27 Nov 2018.
- Clary CM, Kestens Y. Field validation of secondary data sources: a novel measure of representativity applied to a Canadian food outlet database. *Int J Behav Nutr Phys Act*. 2013;10:77.

35. D'Angelo H, Fleischhacker S, Rose SW, Ribisl KM. Field validation of secondary data sources for enumerating retail tobacco outlets in a state without tobacco outlet licensing. *Health Place*. 2014;28:38–44.
36. R: a language and environment for statistical computing, Vienna, Austria. <https://www.R-project.org/>.
37. osmar: OpenStreetMap and R. *R Journal*. 2012. <http://osmar.r-forge.r-project.org/Rjpreprint.pdf>. Accepted for publication on 2012-08-14.
38. Overpass turbo. <https://overpass-turbo.eu/>. Accessed 17 Jan 2019.
39. Lampros Mouselimis. geojsonR: a GeoJson processing toolkit. R package version 1.0.0. 2017. <https://CRAN.R-project.org/package=geojsonR>. Accessed 11 Sept 2018.
40. Jeroen Ooms. The jsonlite package: a practical and consistent mapping between JSON data and R objects. 2014. [arXiv:1403.2805](https://arxiv.org/abs/1403.2805) [stat.CO]. <https://arxiv.org/abs/1403.2805>.
41. Bivand R, Keitt T, Rowlingson B. rgdal: bindings for the Geospatial Data Abstraction Library. R package version 1.2-4. 2016.
42. Kahle D, Wickham H. ggmap: spatial visualization with ggplot2. *R J*. 2013;5(1):144–61.
43. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2009.
44. Schneider S, Gruber J. Neighbourhood deprivation and outlet density for tobacco, alcohol and fast food: first hints of obesogenic and addictive environments in Germany. *Public Health Nutr*. 2013;16(7):1168–77.
45. Economos CD, Hatfield DP, King AC, Ayala GX, Pentz MA. Food and physical activity environments: an energy balance approach for research and practice. *Am J Prev Med*. 2015;48(5):620–9.
46. Ludwig DS, Ebbeling CB. The carbohydrate-insulin model of obesity: beyond "Calories In, Calories Out". *JAMA Intern Med*. 2018;178(8):1098–103.
47. Hall KD, Ayuketah A, Brychta R, et al. Ultra-processed diets cause excess calorie intake and weight gain: an inpatient randomized controlled trial of ad libitum food intake. *Cell Metab*. 2019. <https://doi.org/10.1016/j.cmet.2019.05.008>.
48. Congdon P. Variations in obesity rates between US counties: impacts of activity access, food environments, and settlement patterns. *Int J Environ Res Public Health*. 2017 Sep 7;14(9):1023
49. Diez J, Bilal U, Cebrecos A, Buczynski A, Lawrence RS, Glass T, Escobar F, Gittelsohn J, Franco M. Understanding differences in the local food environment across countries: a case study in Madrid (Spain) and Baltimore (USA). *Prev Med*. 2016;89:237–44.
50. Roshandel S, Zheng Z, Washington S. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. *Accid Anal Prev*. 2015;79:198–211.
51. Gebel K, Ding D, Foster C, Bauman AE, Sallis JF. Improving current practice in reviews of the built environment and physical activity. *Sports Med*. 2015;45(3):297–302.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

