Current Epidemiology Reports (2022) 9:175–182
https://doi.org/10.1007/s40471-022-00296-7

Machine Learning Approaches for Measuring Neighborhood Environments in Epidemiologic Studies

Andrew G. Rundle[1] · Michael D. M. Bader[2] · Stephen J. Mooney[3]

Accepted: 3 April 2022 / Published online: 30 June 2022

Abstract

Purpose of review Innovations in information technology, initiatives by local governments to share administrative data, and growing inventories of data available from commercial data aggregators have immensely expanded the information available to describe neighborhood environments, supporting an approach to research we call Urban Health Informatics. This review evaluates the application of machine learning to this new wealth of data for studies of the effects of neighborhood environments on health.

Recent findings Prominent machine learning applications in this field include automated image analysis of archived imagery such as Google Street View images, variable selection methods to identify neighborhood environment factors that predict health outcomes from large pools of exposure variables, and spatial interpolation methods to estimate neighborhood conditions across large geographic areas.

Summary In each domain, we highlight successes and cautions in the application of machine learning, particularly highlighting legal issues in applying machine learning approaches to Google's geo-spatial data.

Keywords Machine learning · Neighborhood environments · Google Street View · Spatial interpolation · Neighborhood-wide association studies · Urban health informatics

Introduction

Neighborhood health research seeks to explain how neighborhood characteristics such as built features, social and economic conditions, and chemical and particulate pollutant concentrations affect residents' health. In addition to public health and medicine, urban sociologists, planners, and architects contribute to the field. The methods used for study design and data analysis draw from sociology and environmental epidemiology. Findings from this body of research have influenced policymakers, architects, planners, and commercial entities, including supporting city policies that encourage health-promoting businesses like grocery stores and that establish urban design, architecture, and planning guidelines [1–3].

Defining and measuring neighborhood features presents challenges for neighborhood health effects research. Physical and social characteristics of neighborhoods vary widely and at multiple geographic scales in ways that make them difficult to characterize [4]. But innovations in information technology, the greater willingness of local governments to share administrative data, and a growing awareness of the types of data that can be purchased from commercial data aggregators have meant that the information available to characterize neighborhoods has expanded immensely over the past 20 years. These data have been linked to health data from surveys, health surveillance systems, schools, medical records, and epidemiologic studies [5–9]. We refer to urban health informatics as the use of information technology to tap into, organize, cross-link, and analyze the massive data stream produced by, and about, urban centers, to understand the

This article is part of the Topical Collection on *Environmental Epidemiology*

✉ Andrew G. Rundle
Agr3@cumc.columbia.edu

[1] Department of Epidemiology, Mailman School of Public Health, Columbia University, New York City, NY, USA

[2] Department of Sociology, Johns Hopkins University, Baltimore, MD, USA

[3] Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA

health of residents [10]. Given the expanding data available for research, the field of neighborhood health research has started to use increasingly complex techniques to process these data more efficiently and accurately to characterize and identify exposures that affect health and health behaviors. Many of these techniques are drawn from machine learning, which we define here as the use of algorithms to uncover patterns in data that are then used without human intervention to make predictions about other data.

This review describes and evaluates three domains of urban health informatics in which innovative machine learning approaches have recently been applied (see Table 1). The first is automated image analysis of archived imagery such as Google Street View images. The second involves variable selection methods to identify neighborhood environment factors that predict health outcomes from large pools of exposure data with candidate variables. The third application uses spatial interpolation methods to estimate neighborhood conditions across large areas, using data collected at a limited number of sites.

Automated Image Analysis of Archived Imagery

The basis for much of the research on neighborhood health effects is derived from a method known as systematic social observation (SSO), also called neighborhood auditing [11, 12]. The method involves developing standard protocols to evaluate physical and social conditions (e.g., abandoned buildings, graffiti, and trash on the streets) on a systematic sample of locations (often street blocks or intersections) [11, 12]. The method was originally implemented by trained auditors visiting locations in person. In some cases, this involved auditing study participants' neighborhoods when team members visited their homes to conduct face-to-face interviews or collect environmental samples, and in others, it involved auditing locations selected by the researchers to ensure geographic coverage of an area of research interest [11, 13].

SSO audit instruments have been developed and validated for measuring pedestrian safety features, pedestrian infrastructure, neighborhood physical disorder, advertising for alcohol and tobacco, and food environments. However, the in-person SSO approach has several limitations, including the time and expense for researchers to travel to the sampled locations, ensuring the physical safety of auditors, and community acceptance of researchers observing neighborhoods [14].

To address these problems, researchers developed virtual systematic social observations (VSSOs) [14, 15]. Instead of sending raters to physically inspect the block, VSSOs have trained auditors to use Google Street View's archived imagery to collect observational data from streets or intersections [14, 16, 17]. Other implementations of VSSOs include collecting data from archives of images from public webcams [18]. Several existing in-person SSO protocols have been adapted for use in the virtual environment, and web-based tools have been developed to manage VSSO studies, notably the Computer-Aided Neighborhood Visual Assessment System (CANVAS) [19]. This approach has been shown to require much less time and expense than in-person audits while equivalently measuring the physical environment and offers the possibility of rating much larger geographic expanses, including national samples [16, 17, 19]. The strengths and weaknesses of VSSO have been previously discussed extensively [14–16].

Several groups have sought to expand the VSSO approach by replacing trained human observers with machine learning tools that automatically identify features in downloaded Google Street View images (see Fig. 1). This approach typically requires mass downloading of Street View images in order to efficiently process the images on high-performance computational clusters. There have been notable successes in training machines to quantify trees and green space and to identify traffic control signs, crosswalks, single-lane roads, and utility wires in Street View panoramas [20–26]. Recent work has also used health data from surveys to train a generative adversarial networks (GAN) machine learning model to identify health-related features in Street View images

Table 1 Summary of machine learning applications

Application	Uses	Issues
Automated image analysis	Accelerate virtual systematic social observation (VSSO) methods. Increase the density of sampled locations and expand the geographic coverage of VSSO studies	Legal issues with Google's terms of use for Google Maps and Street View
Variable selection	Application of GWAS and EWAS-style studies to large pools of neighborhood-level variables	Choosing from the array of possible selection algorithms: results from different algorithms may disagree
Spatial interpolation	Characterize neighborhood conditions over large areas using data from a sample of locations	Researchers using "off-the-shelf" existing data versus researchers setting their own sampling plan for collecting data: need to account for uncertainty in the estimation of the neighborhood-level data

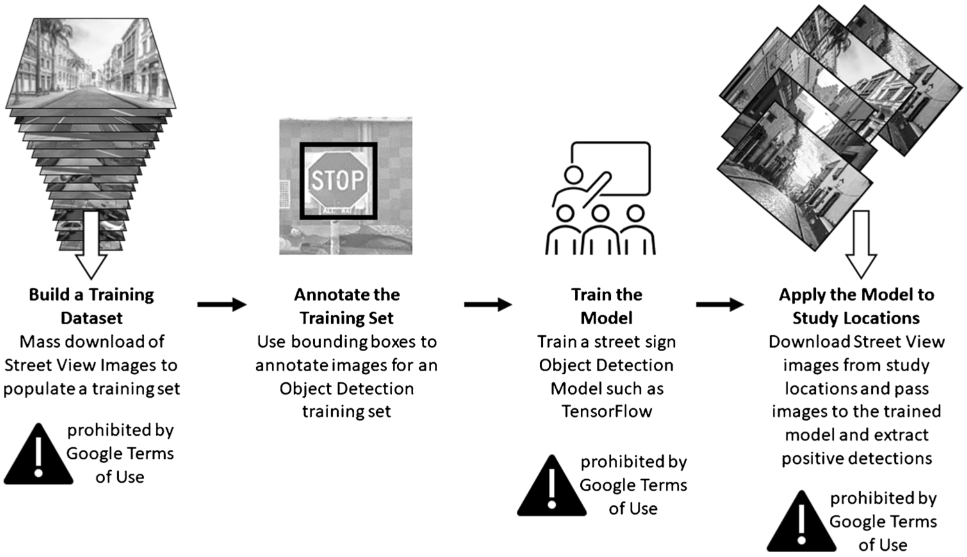
STOP
Build a Training Dataset
Mass download of Street View Images to populate a training set
prohibited by Google Terms of Use
Annotate the Training Set
Use bounding boxes to annotate images for an Object Detection training set
Train the Model
Train a street sign Object Detection Model such as TensorFlow
prohibited by Google Terms of Use
Apply the Model to Study Locations
Download Street View images from study locations and pass images to the trained model and extract positive detections
prohibited by Google Terms of Use

Fig. 1 A common workflow for machine learning applied to Google Street View Images and its relationship to activities prohibited by Google's terms of use [26]

[27]. They found that respondents' self-reported physical function using the PF-10 was associated with urban greenery, including tree height and building height identified in the images by the GAN model [27]. Another approach has been to have raters provide an evaluation of a single dimension, for example, safety, and then use the raters' ratings to train machine algorithms to evaluate safety on other blocks [28]. Machine-learning-based approaches to identifying and quantifying key features of the built environment from Street View imagery have the potential to fully automate the conduct of VSSO and radically speed up the creation of data on neighborhood conditions.

We caution that machine learning approaches may ultimately incur legal challenges. As of January 2022—and since at least February 2018—Google's overall terms of use and those of their Google Maps product (including "Geo Guidelines" included in the Google Maps terms) prohibit downloading and storing imagery, recreating panoramic views from downloaded image tiles, and the use of "… applications to analyze and extract information from the Street View imagery" (see Fig. 1). [29, 30] The creation of measures of trees from Google Maps products is specifically given as an example of prohibited uses [29]. The terms of use also prohibit the use of Google Maps-derived data in point-in-polygon analyses, a geographic information systems technique commonly used in neighborhood health effects research. The Geo Guidelines explicitly state that nonprofit and academic uses are not exempt from the terms: "these restrictions apply to all academic, nonprofit, and commercial projects," and further that they will not grant exceptions: "If your use is not allowed, we are not able to grant exceptions, so please do not submit a request."[30] These prohibitions are not based on copyright, for which researchers might invoke the concept of fair use, but based on the contract a user agrees to follow (an "end user's licensing agreement") by accessing Google Maps. [31] The enforceability of such contracts is an area of active litigation and, as such, is unclear; researchers who conduct this type of research and the journals that publish the resulting papers incur some legal risk so long as the law remains unsettled [8, 32].

Variable Selection Approaches to Analyzing "Big Data"

A second area of innovation regarding machine learning and neighborhood health research involves variable selection in prediction models using "hypothesis-free" analyses. Vast amounts of data can be processed and linked together using geographic information systems (GIS), including census data, business listing data (e.g., the National Establishment Times Series), social media data, online search data, and administrative data from local, state, and federal governments (e.g., tax, licensing, inspection, maintenance, and enforcement data) to measure neighborhood conditions [33–35]. When all of these data sources are compiled together, the GIS becomes a high-throughput data-generating platform that can produce thousands of variables describing the environment of a neighborhood.

While the quantity and richness of these data are very attractive to researchers, the variables are often highly correlated, causing multicollinearity issues in multivariable data analyses that limit the ability to isolate the effects of individual variables using hypothesis-based approaches to data analysis (see Fig. 2). As in other fields that have faced these issues, like with the -omic arrays used to study genomics and proteomics, and studies of exposures to complex chemical mixtures, researchers studying neighborhood health effects

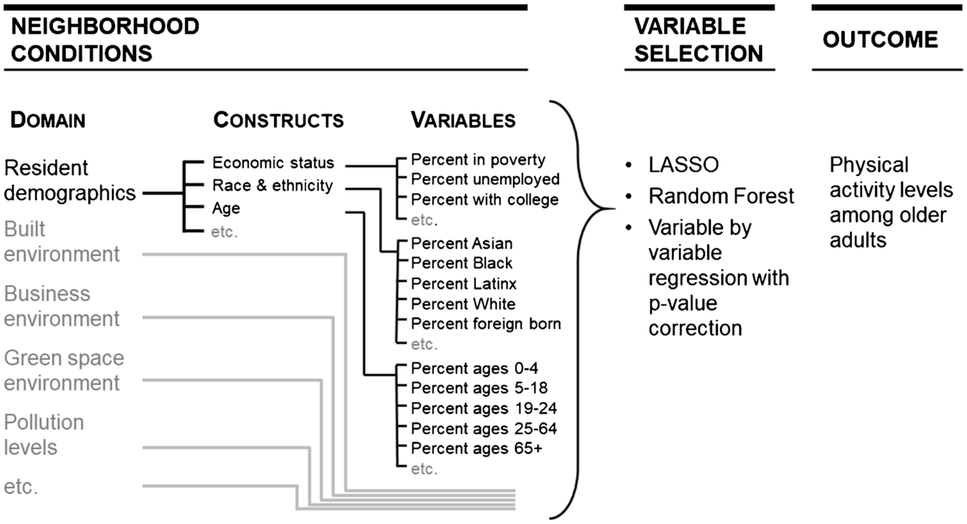

Fig. 2 Schematic diagram of a neighborhood environment-wide association study as applied to selecting variables that predict physical activity levels in older adults [41]

have turned to machine learning and variable selection approaches to identify neighborhood environment variables associated with health outcomes [36–38]. Using genome-wide association studies (GWAS) [39] and environment-wide association studies (EWAS) [40] as a template, neighborhood health researchers have begun implementing neighborhood environment-wide association studies (NE-WAS) [41–44]. These studies use computer algorithms to identify neighborhood environment variables most strongly associated with health outcomes of interest [41]. Examples of this are efforts to identify neighborhood-level variables that predict prostate cancer aggressiveness, physical activity levels among older adults, COVID-19 mortality, neighborhood-based walking, and violent crime [41, 42, 44–46].

How best to conduct NE-WAS analyses remains unclear. For any of the '-WAS approaches, including NE-WAS, there are numerous algorithms for simultaneously analyzing large quantities of predictor variables, and researchers debate which of these algorithms identifies relationships of greatest scientific interest [36–38, 47–50]. Simple approaches include calculating an odds ratio and *p*-value for each predictor variable independently and applying a multiple comparisons correction to the threshold for declaring statistical significance [41, 42]. Other approaches, which are vastly more computationally complex, include reducing the number of dimensions (e.g., through principal components, latent classes, *k*-means clustering, etc.) and variable selection steps (e.g., LASSO or stepwise regression) [38, 47, 48, 50]. These approaches often sub-set the full dataset into training and validation subsets and then use cross-validation to select tuning parameters from the training subset, validate the parameters on the validation subset, and then apply the parameters to the entire dataset [38, 47, 48, 50].

These complex analytic techniques, developed by computer data scientists, make principled use of patterns in the data to build robust classification and prediction models. However, their successful use in neighborhood research has been more limited. One reason for this is that algorithms designed to select the variables that together best predict an outcome do not, in general, select a subset of distinct variables that are most causally relevant [51]. Moreover, because neighborhood predictors are frequently strongly correlated with each other (e.g., % of households living in poverty is highly correlated with median household income), naive application of variable selection algorithms (e.g., with no pre-selection of variables to ensure measures cover only a subset of domains) results in highly unstable selections—minor tweaks to the dataset result in a very different selection of variables [52].

To address a similar issue of highly correlated predictor variables in GWAS studies, an initial process of "pruning" is often used to remove from a dataset the data for one or the other of two genetic loci with SNPs that show high linkage disequilibrium (high correlation) [53–55]. Typically, in pruning GWAS data sets, the data for the loci with the highest minor allele frequency (MAF) is kept in the dataset, and the loci with the lowest MAF are removed from the dataset [53–55]. Because neighborhood-level variables are commonly continuously distributed, there is not an exact analogy to MAF to guide pruning decisions in NE-WAS studies. However, considerations related to the effects of measurement error on bias in neighborhood health effects studies can be used to guide decisions regarding which variables to prune. Measurement error in underlying data (e.g., personal income data collected in the American Community Survey) that are aggregated up to create neighborhood-level measures (e.g., % of the population living in poverty or median household income or per capita income) causes systematic bias away from the null in neighborhood health effects studies when the neighborhood level variable is expressed as a proportion (e.g., % of the population living in poverty) but not for variables expressed a continuous scales (e.g., median household income) [56]. Thus, for variable pruning in NE-WAS studies, within sets of highly correlated variables, we recommend removing variables expressed as proportions.

Spatial Interpolation to Estimate Neighborhood Conditions

Spatial interpolation is another area where machine learning approaches have been applied to neighborhood health research, including in studies of air pollution, of neighborhood physical disorders, and sidewalk conditions [4, 57–59]. Spatial interpolation models such as kriging and land use regression estimate neighborhood conditions at all locations across a geographic area using measured data from only a sample of locations in the region [4, 57, 60–63]. The estimates are based on the measured values at the sampled locations, distances to sampled points, and the spatial correlation between measured values at sampled points. The estimation models often also include external data measured at all locations in the target geographic area (e.g., home prices or distance to roads). Spatial interpolation can use machine learning techniques, including automated variable selection techniques that can select predictors and functional forms included in a final model from a class of candidate predictors and leave-one-out cross-validation that "tunes" the interpolation model [64].

Spatial interpolation can be implemented using several approaches. One such approach is land use regression, which is commonly used in air pollution studies. This method uses data on relevant land use features (e.g., industrial zoning, road density, vehicle traffic, and pollution point sources) to create a regression model predicting pollution levels measured at air monitoring stations [60, 61, 65]. This regression model is then used to estimate air pollution at all other locations in the target area [60].

Another spatial interpolation approach is ordinary kriging, which uses the spatial correlation between the variable values at each sampled location (e.g., particulate matter measured at air monitors) to estimate values at all non-sampled locations in the geographic area of interest [4, 57]. Universal kriging is an extension of ordinary kriging that uses additional external data that can be measured at all locations in the geographic area to supplement the information represented by the spatial correlations [66]. Universal kriging can be viewed as jointly modeling ordinary kriging and land-use regression [60]. In a study of particulate matter air pollution, for example, relevant external data might be vehicle traffic volume. In our work using universal kriging to estimate neighborhood physical disorder in four US cities, we found that using a measure of housing vacancy in universal kriging improved estimation over ordinary kriging for Philadelphia and Detroit. [66] More complex approaches apply automated techniques to a suite of candidate variables to improve these models, either by algorithmically selecting specific environment variables to include in a model or by applying dimensionality reduction techniques to identify underlying factors that maximize the predictive value of external data [64, 67]. These complex approaches are more commonly used to characterize exposures to environmental pollutants than neighborhood built or social environment factors, but they present a promising direction for new methods to describe neighborhood conditions.

There are several considerations to be acknowledged when using spatial interpolation techniques. When data are available from a series of locations "off the shelf" (often the case for air monitoring data but can also be true for other administrative data), the spatial distribution of these locations is often not optimized for making interpolation estimates at non-sampled locations [57]. In these off-the-shelf scenarios, the estimates can have greater uncertainty than would be seen if the sampling were designed to optimize spatial interpolation. In air pollution studies, the air monitoring stations set up for administrative purposes are often located in such a way that there is higher uncertainty for the interpolated estimates near the borders of a city or region [57]. Spatial interpolation works best when the sample of locations where data will be collected is chosen to optimize spatial interpolation algorithms [60, 65]. There are considerations when researchers choose sample locations to collect data from; should the sample be based on population distribution of land area [4]? When chosen sampled locations to reflect population distribution in an area, the resulting sample will on average represent the population but not necessarily geography and land uses [4]. For example, non-inhabited but relevant areas like industrial plants or parks would be excluded from a sample of locations selected based on population distribution.

When interpolated neighborhood exposure values are used in a regression analysis predicting some health outcomes, the uncertainty in the neighborhood measurements should be included in analyses of health outcomes. Not doing so will lead to underestimating the uncertainty of parameter estimates measuring the association between the neighborhood exposure and the health outcome. A solution to this issue uses a framework very similar to multiple imputations for missing data [4, 58]. Instead of a single value being estimated for each neighborhood location of interest, multiple values for each location are estimated from the interpolation model to represent the uncertainty in the estimation process [4, 58]. These multiple exposure data sets are then each analyzed to predict the health outcome and the estimated effect size and its corresponding standard error from each data set are combined to create a pooled estimated effect size and standard error. As in multiple imputations, the pooled standard error from analyzes of multiple estimated data sets better expresses the uncertainty in the observed association between exposure and outcome [68].

Conclusion

Revolutions in information technology, the greater willingness of local governments to share their administrative data, and a growing awareness of the types of data that can be purchased from commercial data aggregators mean that the information available to characterize neighborhoods has expanded immensely over the past 20 years. As data availability has expanded, researchers studying neighborhood health effects have started to utilize machine learning approaches to measure and identify neighborhood features that influence health. Spatial interpolation methods, particularly for estimating air pollution, have the most established track record for the use of machine learning in characterizing neighborhood environments. While GWAS have been employed for over twenty years in genetic epidemiology, similar variable selection approaches have just begun to be implemented with neighborhood-level data. In other fields that grapple with large quantities of intercorrelated predictor variables, such as -omics and the study of chemical mixtures, there is debate over which variable selection algorithms are most appropriate. Lessons learned from these other fields are likely to be applicable to neighborhood-level data.

Of the machine learning approaches that have been used, it is the application of automated image analysis that has perhaps most captured the imagination of researchers. Unfortunately, the terms of use for Google Street View, the data source most commonly used for VSSOs, expressly prohibit the use of machine learning to identify features in Street View images. We hope that ongoing litigation will clarify the enforceability of terms of use and that companies that create these valuable spatial data sets will make them available for health research. Yelp, for example, has a special program for academic researchers who wish to use their data[69]. Until then, journals and Institutional Review Boards should pay attention to the use of data acquired without appropriate licenses or in ways contrary to the terms of use.

Funding Drs. Rundle, Mooney, and Bader were supported by a grant from the National Institute on Alcohol Abuse and Alcoholism (R01AA028552); Dr. Rundle was additionally supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (R01RDK127139) and National Institute of Mental Health (R01MH128734); Dr. Mooney was also supported by a grant from the National Library of Medicine (R00LM012868).

Declarations

Competing Interests The authors declare no competing interests.

Ethics Approval and Consent to Participate This article does not contain any studies with humans or animals performed by any of the authors.

References

1. Lovasi GS, Bader MD, Rundle AG, Neckerman KM. Healthy and Unhealthy Food Sources in NYC: Tracing the generation, evolution, and dissemination of policy-relevant research on the food environment. Case Study 1. In: Hiatt RA, editor. Population Health: The Translation of Research to Policy. New York, NY: Milbank Memorial Fund; 2018.
2. International Well Building Institute: WELL Building and WELL Community Certification. 2017. https://www.wellcertified.com/our-standard Accessed Jan 2022.
3. Lee KK. Developing and implementing the Active Design Guidelines in New York City. Health Place. 2012;18(1):5–7. https://doi.org/10.1016/j.healthplace.2011.09.009.
4. Bader MDM, Ailshire JA. Creating measures of theoretically relevant neighborhood attributes at multiple spatial scales. Sociol Methodol. 2014;44(1):322–68. https://doi.org/10.1177/0081175013516749.
5. Freeman L, Neckerman K, Schwartz-Soicher O, Quinn J, Richards C, Bader MD, et al. Neighborhood walkability and active travel (walking and cycling) in New York City. J Urban Health. 2013;90(4):575–85. https://doi.org/10.1007/s11524-012-9758-7.
6. Tabaei BP, Rundle AG, Wu WY, Horowitz CR, Mayer V, Sheehan DM, et al. Associations of residential socioeconomic, food, and built environments with glycemic control in persons with diabetes in New York City From 2007–2013. Am J Epidemiol. 2018;187(4):736–45. https://doi.org/10.1093/aje/kwx300.
7. Lebwohl B, Genta RM, Kapel RC, Sheehan D, Lerner NS, Green PH, et al. Procedure volume influences adherence to celiac disease guidelines. Eur J Gastroenterol Hepatol. 2013;25(11):1273–8. https://doi.org/10.1097/MEG.0b013e3283643542.
8. HIQ Labs, Inc v. LINKEDIN Corporation, (2019).
9. Lovasi GS, Quinn JW, Rauh VA, Perera FP, Andrews HF, Garfinkel R, et al. Chlorpyrifos exposure and urban residential environment characteristics as determinants of early childhood neurodevelopment. Am J Public Health. 2011;101(1):63–70. https://doi.org/10.2105/AJPH.2009.168419.
10. Rundle AG. Built Environment and Health (BEH) Research Group, About. 2021. https://beh.columbia.edu/about-2/. Accessed Jan 2022.

11. Raudenbush SW, Sampson RJ. Ecometrics: toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. 1999;29(1):1–41. https://doi.org/10.1111/0081-1750.00059.
12. Sampson RJ, Raudenbush SW. Seeing disorder: neighborhood stigma and the social construction of "broken windows". 2004;67(4):319–42. doi:https://doi.org/10.1177/019027250406700401.
13. Fragile Families and Child Wellbeing Study: Data Contents and Overview. 2022. https://fragilefamilies.princeton.edu/data-and-documentation/data-contents-overview. Accessed Jan 2022.
14. Rundle AG, Bader MD, Richards CA, Neckerman KM, Teitler JO. Using Google Street View to audit neighborhood environments. Am J Prev Med. 2011;40(1):94–100. https://doi.org/10.1016/j.amepre.2010.09.034.
15. Bader MDM, Mooney SJ, Bennett B, Rundle AG. The promise, practicalities, and perils of virtually auditing neighborhoods using Google Street View. 2017;669(1):18–40. https://doi.org/10.1177/0002716216681488.
16. Mooney SJ, Bader MDM, Lovasi GS, Teitler JO, Koenen KC, Aiello AE, et al. Street audits to measure neighborhood disorder: virtual or in-person? Am J Epidemiol. 2017;186(3):265–73. https://doi.org/10.1093/aje/kwx004.
17. Mooney SJ, DiMaggio CJ, Lovasi GS, Neckerman KM, Bader MD, Teitler JO, et al. Use of Google Street View to assess environmental contributions to pedestrian injury. Am J Public Health. 2016;106(3):462–9. https://doi.org/10.2105/AJPH.2015.302978.
18. Hipp JA, Adlakha D, Eyler AA, Chang B, Pless R. Emerging technologies: webcams and crowd-sourcing to identify active transportation. Am J Prev Med. 2013;44(1):96–7. https://doi.org/10.1016/j.amepre.2012.09.051.
19. Bader MD, Mooney SJ, Lee YJ, Sheehan D, Neckerman KM, Rundle AG, et al. Development and deployment of the computer assisted neighborhood visual assessment system (CANVAS) to measure health-related neighborhood conditions. Health Place. 2015;31:163–72. https://doi.org/10.1016/j.healthplace.2014.10.012.
20. Nguyen QC, Keralis JM, Dwivedi P, Ng AE, Javanmardi M, Khanna S, et al. Leveraging 31 million Google Street View images to characterize built environments and examine county health outcomes. Public Health Rep. 2021;136(2):201–11. https://doi.org/10.1177/0033354920968799.
21. Larkin A, Hystad P. Evaluating street view exposure measures of visible green space for health research. J Expo Sci Environ Epidemiol. 2019;29(4):447–56. https://doi.org/10.1038/s41370-018-0017-1.
22. Mennis J, Li X, Meenar M, Coatsworth JD, McKeon TP, Mason MJ. Residential greenspace and urban adolescent substance use: exploring interactive effects with peer network health, sex, and executive function. Int J Environ Res Public Health. 2021;18(4). doi:https://doi.org/10.3390/ijerph18041611.
23. Jodas DS, Yojo T, Brazolin S, Velasco GDN, Papa JP. Detection of trees on street-view images using a convolutional neural network. Int J Neural Syst. 2022;32(1):2150042. https://doi.org/10.1142/S0129065721500428.
24. Thirlwell A, Arandjelovic O. Big data driven detection of trees in suburban scenes using visual spectrum eye level photography. Sensors (Basel). 2020;20(11). doi:https://doi.org/10.3390/s20113051.
25. Lu Y. The association of urban greenness and walking behavior: using Google Street View and deep learning techniques to estimate residents' exposure to urban greenness. Int J Environ Res Public Health. 2018;15(8). doi:https://doi.org/10.3390/ijerph15081576.
26. Campbell A, Both A, Sun Q. Detecting and mapping traffic signs from Google Street View images using deep learning and GIS. Computers, Environment and Urban Systems. 2019;77:101350.: https://doi.org/10.1016/j.compenvurbsys.2019.101350.
27. Rachele JN, Wang J, Wijnands JS, Zhao H, Bentley R, Stevenson M. Using machine learning to examine associations between the built environment and physical function: a feasibility study. Health Place. 2021;70: 102601. https://doi.org/10.1016/j.healthplace.2021.102601.
28. Naik N, Philipoom J, Raskar R, Hidalgo C. Streetscore -- predicting the perceived safety of one million streetscapes. IEEE Conference on Computer Vision and Pattern Recognition Workshops 2014. p. 793–9.
29. Google: Google Maps Platform Terms of Service. 2020. https://cloud.google.com/maps-platform/terms. Accessed Jan 2022.
30. Google: Google Maps, Google Earth, and Street View. 2020. https://about.google/brand-resource-center/products-and-services/geo-guidelines/#street-view. Accessed Jan 2022.
31. Google: Google Maps APIs Terms of Service. 2018. https://developers.google.com/maps/terms-20180207?_ga=2.84925724.401285425.1641176208-1911476959.1641176208. Accessed Jan 2022.
32. Stringam B, Gerdes JH, Anderson CK. Legal and ethical issues of collecting and using online hospitality data.0(0):19389655211040434. https://doi.org/10.1177/19389655211040434.
33. Rundle A, Rauh VA, Quinn J, Lovasi G, Trasande L, Susser E, et al. Use of community-level data in the National Children's Study to establish the representativeness of segment selection in the Queens Vanguard Site. Int J Health Geogr. 2012;11:18. https://doi.org/10.1186/1476-072X-11-18.
34. Hirsch JA, Moore KA, Cahill J, Quinn J, Zhao Y, Bayer FJ, et al. Business data categorization and refinement for application in longitudinal neighborhood health research: a methodology. J Urban Health. 2021;98(2):271–84. https://doi.org/10.1007/s11524-020-00482-2.
35. Laszkowska M, Shiwani H, Belluz J, Ludvigsson JF, Green PHR, Sheehan D, et al. Socioeconomic vs health-related factors associated with google searches for gluten-free diet. Clin Gastroenterol Hepatol. 2018;16(2):295–7. https://doi.org/10.1016/j.cgh.2017.07.042.
36. Czarnota J, Gennings C, Wheeler DC. Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk. Cancer Inform. 2015;14(Suppl 2):159–71. https://doi.org/10.4137/CIN.S17295.
37. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. J Agric Biol Environ Stat. 2015;20(1):100–20. https://doi.org/10.1007/s13253-014-0180-3.
38. Taylor KW, Joubert BR, Braun JM, Dilworth C, Gennings C, Hauser R, et al. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environ Health Perspect. 2016;124(12):A227–9. https://doi.org/10.1289/EHP547.
39. Neale BM, Purcell S. The positives, protocols, and perils of genome-wide association. Am J Med Genet B Neuropsychiatr Genet. 2008;147B(7):1288–94. https://doi.org/10.1002/ajmg.b.30747.
40. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS ONE. 2010;5(5): e10746. https://doi.org/10.1371/journal.pone.0010746.
41. Mooney SJ, Joshi S, Cerda M, Kennedy GJ, Beard JR, Rundle AG. Contextual correlates of physical activity among older adults: a neighborhood environment-wide association study (NE-WAS). Cancer Epidemiol Biomarkers Prev. 2017;26(4):495–504. https://doi.org/10.1158/1055-9965.EPI-16-0827.

42. Lynch SM, Mitra N, Ross M, Newcomb C, Dailey K, Jackson T, et al. A neighborhood-wide association study (NWAS): example of prostate cancer aggressiveness. PLoS ONE. 2017;12(3): e0174548. https://doi.org/10.1371/journal.pone.0174548.
43. Hu H, Zhao J, Savitz DA, Prosperi M, Zheng Y, Pearson TA. An external exposome-wide association study of hypertensive disorders of pregnancy. Environ Int. 2020;141: 105797. https://doi.org/10.1016/j.envint.2020.105797.
44. Hu H, Zheng Y, Wen X, Smith SS, Nizomov J, Fishe J, et al. An external exposome-wide association study of COVID-19 mortality in the United States. Sci Total Environ. 2021;768: 144832. https://doi.org/10.1016/j.scitotenv.2020.144832.
45. Mooney SJ, Hurvitz PM, Moudon AV, Zhou C, Dalmat R, Saelens BE. Residential neighborhood features associated with objectively measured walking near home: revisiting walkability using the automatic context measurement tool (ACMT). Health Place. 2020;63: 102332. https://doi.org/10.1016/j.healthplace.2020.102332.
46. Redfern J, Sidorov K, Rosin PL, Corcoran P, Moore SC, Marshall D. Association of violence with urban points of interest. PLoS ONE. 2020;15(9): e0239840. https://doi.org/10.1371/journal.pone.0239840.
47. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–32. https://doi.org/10.1038/nrg3920.
48. Tan MS, Cheah PL, Chin AV, Looi LM, Chang SW. A review on omics-based biomarkers discovery for Alzheimer's disease from the bioinformatics perspectives: statistical approach vs machine learning approach. Comput Biol Med. 2021;139: 104947. https://doi.org/10.1016/j.compbiomed.2021.104947.
49. Kino S, Hsu YT, Shiba K, Chien YS, Mita C, Kawachi I, et al. A scoping review on the use of machine learning in research on social determinants of health: trends and research prospects. SSM Popul Health. 2021;15: 100836. https://doi.org/10.1016/j.ssmph.2021.100836.
50. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article25. https://doi.org/10.2202/1544-6115.1309.
51. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. Stat Methods Med Res. 2012;21(1):7–30. https://doi.org/10.1177/0962280210387717.
52. Mooney S. The impact of built and social environment on physical activity among older adults. New York, NY: Columbia University; 2016.
53. Prive F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018;34(16):2781–7. https://doi.org/10.1093/bioinformatics/bty185.
54. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27(2): e1608. https://doi.org/10.1002/mpr.1608.
55. Calus MPL, Vandenplas J. SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. Genet Sel Evol. 2018;50(1):34. https://doi.org/10.1186/s12711-018-0404-z.
56. Mooney SJ, Richards CA, Rundle AG. There goes the neighborhood effect: bias owing to nondifferential measurement error in the construction of neighborhood contextual measures. Epidemiology. 2014;25(4):528–35. https://doi.org/10.1097/EDE.0000000000000113.
57. Jerrett M, Burnett RT, Ma R, Pope CA 3rd, Krewski D, Newbold KB, et al. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology. 2005;16(6):727–36. https://doi.org/10.1097/01.ede.0000181630.15826.7d.
58. Mooney SJ, Bader MD, Lovasi GS, Neckerman KM, Teitler JO, Rundle AG. Validity of an ecometric neighborhood physical disorder measure constructed by virtual street audit. Am J Epidemiol. 2014;180(6):626–35. https://doi.org/10.1093/aje/kwu180.
59. Plascak JJ, Llanos AAM, Chavali LB, Xing CY, Shah NN, Stroup AM, et al. Sidewalk conditions in Northern New Jersey: using Google Street View imagery and ordinary kriging to assess infrastructure for walking. Prev Chronic Dis. 2019;16:E60. https://doi.org/10.5888/pcd16.180480.
60. Clougherty JE, Kheirbek I, Eisl HM, Ross Z, Pezeshki G, Gorczynski JE, et al. Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the New York City Community Air Survey (NYCCAS). J Expo Sci Environ Epidemiol. 2013;23(3):232–40. https://doi.org/10.1038/jes.2012.125.
61. Ross Z, Ito K, Johnson S, Yee M, Pezeshki G, Clougherty JE, et al. Spatial and temporal estimation of air pollutants in New York City: exposure assignment for use in a birth outcomes study. Environ Health. 2013;12:51. https://doi.org/10.1186/1476-069X-12-51.
62. Cressie N. Statistics for spatial data. Revised Edition. Wiley Series in Probability and Statistics. New York, NY: Wiley; 1993.
63. Isaaks E, Srivastava R. An introduction to applied geostatistics. New York NY: Oxford University Press; 1989.
64. Couckuyt I, Forrester A, Gorissen D, De Turck F, Dhaene T. Blind kriging: implementation and performance analysis. Adv Eng Softw. 2012;49:1–13. https://doi.org/10.1016/j.advengsoft.2012.03.002.
65. Matte TD, Ross Z, Kheirbek I, Eisl H, Johnson S, Gorczynski JE, et al. Monitoring intraurban spatial patterns of multiple combustion air pollutants in New York City: design and implementation. J Expo Sci Environ Epidemiol. 2013;23(3):223–31. https://doi.org/10.1038/jes.2012.126.
66. Mooney SJ, Bader MD, Lovasi GS, Neckerman KM, Rundle AG, Teitler JO. Using universal kriging to improve neighborhood physical disorder measurement. Sociol Methods Res. 2020;49(4):1163–85. https://doi.org/10.1177/0049124118769103.
67. Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV et al. A regionalized national universal kriging model using partial least squares regression for estimating annual PM2.5 concentrations in epidemiology. Atmos Environ (1994). 2013;75:383–92. https://doi.org/10.1016/j.atmosenv.2013.04.015.
68. Rubin D. Mulitple Imputation for Nonresponse in Surveys. Wiley Classics Library. Hoboken, NJ: Wiley-Interscience; 2004.
69. Yelp: Yelp Open Dataset. 2022. https://www.yelp.com/dataset. Accessed Jan 2022.