Remote Sensing Applications: Society and Environment 30 (2023) 100950

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

journal homepage: www.elsevier.com/locate/rsase

Identifying characteristics of the natural and built environment associated with child development: A pilot study integrating google street view, computer vision models, and bioinformatic approaches

Lee E. Voth-Gaeddert [a, b, *], Xanthe Hunt [c], Mark Tomlinson [c, d], Alastair Van Heerden [a, e]

[a] SAMRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, South Africa
[b] Biodesign Center for Health Through Microbiomes, Arizona State University, Tempe, AZ, USA
[c] Institute for Life Course Health Research, Department of Global Health, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa
[d] School of Nursing and Midwifery, Queens University, Belfast, UK
[e] Center for Community Based Research, Human Sciences Research Council, Pietermaritzburg, South Africa

ARTICLE INFO

Keywords:
GSV images
Image classification
Child development
Environmental exposures
South Africa

ABSTRACT

Recent studies have demonstrated the potential for leveraging computer vision models and Google Street View (GSV) images to identify associations between infrastructure attributes and population health outcomes. However, these studies underutilize the potential available data by focusing on a small set of predetermined indicators. In this study, we integrate these methods and bioinformatic approaches to fundamentally reframe how the complex natural and built environments are represented and evaluated. We demonstrate this integrated approach by assessing significant differences in attributes of the environment surrounding childcare facilities that reported high and low child development outcomes in South Africa. Using a cross-sectional study design, a standard geofence (1.6 km square) was applied to a set of GPS coordinates from N = 86 childcare facilities in South Africa reporting mean child development outcomes in the top and bottom 10% of all facilities surveyed (n = 43 each). GSV images in the geofenced site around the childcare facilities were downloaded via the GSV API. Next, Google's Vision API – a set of algorithms that can generate >10,000 unique labels – was applied to each image, generating a set of 30–50 labels describing features of each image with a minimum accuracy of 60%. Abundances of labels from each site hosting a childcare facility were estimated and normalized. Differences in the abundance values were compared between high and low scoring sites using bar plots, alpha and beta diversity, ordination plots, and Linear discriminant analysis Effect Size (LEfSe). The results suggested that higher abundances of labels associated with transportation and residential buildings or community spaces were present in high scoring sites, while in low scoring sites, labels associated with roads, dry and rural land, and electrical public utilities were significantly more abundant. This exploratory approach can provide a globally scalable method that can generate insights at a granular level for environmental effects on child health.

* Corresponding author. SAMRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, South Africa.
E-mail address: lee.voth-gaeddert@wits.ac.za (L.E. Voth-Gaeddert).

https://doi.org/10.1016/j.rsase.2023.100950
Received 18 October 2022; Received in revised form 6 February 2023; Accepted 4 March 2023
Available online 6 March 2023

1. Introduction

Google Street View (GSV) images have become an increasingly popular data source to evaluate environmental health characteristics of the natural and built environment and their association with various population health outcomes. In addition, recent work has leveraged computer vision models to automate and standardize approaches to identifying or rating environmental health characteristics of interest within these images (Gebru et al., 2017; Nguyen et al., 2019). However, a central challenge in these GSV methods is selecting which inputs or 'exposures' to evaluate (via human identification or algorithmic identification). Previous approaches relied on individual, *a priori* hypotheses and traditional statistical methods to evaluate associations between a single or small set of predictors (e.g., green space, walkability) and the outcome (e.g., safety, chronic disease). GSV has been used to evaluate predictors such as pedestrian safety, walkability and bike-ability, motorized traffic levels (Rundle et al., 2011), neighborhood demographics (Gebru et al., 2017), infrastructure decay and disorder (Rundle et al., 2011), and green spaces (Lu, 2019). These are then evaluated for significant associations with health outcomes such as COVID-19 cases (Nguyen et al., 2020), mental health metrics (Odgers et al., 2012), chronic disease, and mortality (Lu, 2019), among others. The primary benefit of these recent GSV applications is the replacement of the human presence (and bias) in collecting the observational input data. However, the traditional analytical methods applied after data collection is complete, underutilizes the full potential of the 'big data' nature of GSV and computer vision model applications, especially within environmental exposure and human health.

Timing, frequency, diversity, and abundances of different human exposures to environmental pollution sources create a complex problem and identifying a single or small set of factors associated with negative human health outcomes have limited utility. However, given the advancements in the diversity and specificity of computer vision models and availability of GSV images, new methods can be applied that can better represent the complexity of the human-environment interaction in a neighborhood. Specifically, molecular bioinformatic techniques have been developed to evaluate complex microbial communities (such as the human gut microbiome) and their association with health outcomes using a series of metrics and approaches including alpha and beta diversity, ordination plots, and Linear discriminant analysis Effect Size (LEfSe) (Allaband et al., 2019; Dave et al., 2012; Voth-Gaeddert et al., 2019). Neighborhoods or communities can have a similar level of complexity in the local environmental health characteristics (LEHC) present as compared to the human gut microbiome. This similarity allows for a unique application of bioinformatic tools, and a fundamental reframing of how analytical methods (in this case bioinformatic methods) are applied within the GSV-health problem space. This includes how these methods represent the input variables, control for externalities, mitigate biases, and evaluate the effects of the input variables on outcomes of interest.

To demonstrate this potential, we focus on a central human health outcome within South Africa, and globally, namely, child development. A growing body of research confirms the important role of the early years of life for later-life success. Improving child development outcomes such as cognitive and emotional development, are associated with both short-term and long-term physical and mental health challenges (Koletzko et al., 2019; Russell et al., 2022). In addition, child development outcomes are central to many of the targets set in the Sustainable Development Goals (Britto, 2015). Within South Africa, early childhood development continues to improve at a national level, however, the effects of apartheid and post-apartheid policies and programs result in heterogeneity in these improvements, therefore warranting new approaches to monitoring and supporting equitable progress in child development (Barbarin and Richter, 2001; Lu and Treiman, 2011). However, costs to implement robust and sustainable monitoring and reporting systems in South Africa and globally, especially in rural and poor areas, are high. Therefore, new scalable methods, that are easy to contextualize to different settings are required to measure child development, globally. In addition, a significant amount of work has been dedicated to evaluating the association between environmental living conditions (exposures, safety, etc.) and child development outcomes (Walker et al., 2011). Observational studies and environmental sampling have traditionally been used to characterize the environment (Capone et al., 2019; Pickering et al., 2012; Voth-Gaeddert et al., 2018a, 2018b), however, recent advancements in digital technologies have led to an expanding approach to "big data" collection including satellite data and drone imagery (Andres et al., 2018), call detail records (Jones et al., 2018), and integrated weather and topography data (Jeon et al., 2019), among others (Andres et al., 2018). GSV images have demonstrated the potential for an image-based approach to characterizing environments at a granular level, while being scalable globally (Rzotkiewicz et al., 2018). Rzotkiewicz and colleagues conducted a systematic review of GSV in health research and highlighted strengths such as low cost, ease of use, and time savings (compared to physical audits or observations) while noting weaknesses including image resolution, and spatial and temporal availability biased towards wealthier communities and higher population density areas (Rzotkiewicz et al., 2018).

The effects of local environmental health characteristics (LEHCs) on child development, as well as other child health outcomes, are also driven by the LEHCs' diversity and abundance in the setting. Given this, as well as the diversity of trained computer vision algorithms able to identify large numbers of LEHCs, bioinformatic methods can provide an approach to evaluate the diversity and abundance of LECHs and their association with child development. This integration of GSV, computer vision models, and bioinformatics methods could help 1) understand what environmental factors are associated with child development outcomes to improve targeting of programs or policies and 2) predict which locations may have lower child development outcomes based on LEHCs in locations where it is difficult or expensive to have field teams on the ground.

In this study we aim to use GSV images and computer vision models to generate abundance estimates of LEHC labels in sites that host childcare facilities. Bioinformatic approaches will be used to evaluate significant differences in LEHC label abundances between childcare facilities scoring in the top and bottom 10% of a national evaluation of child development among childcare facilities in South Africa.

2. Methods

2.1. Setting

As of 2021, South Africa has a population of 1.3 million children between 4 and 5 years of age (common ages of children in childcare facilities). According to the Thrive by Five Index study, 72% of children 4–5 years of age attend some type of Early Learning Program (i.e., childcare facility) in South Africa, with many more informal childcare arrangements (Giese et al., 2022). Furthermore, 55.3% of children 4–5 years of age fell behind in early learning (as defined by the Thrive by Five study). Statistics South Africa report that in 2020, 51% of children (0–17 years of age) live in poverty (less than $39 or R647 per person per month) (Statistics South Africa, 2020), while malnutrition is high, with 25.1% of children 4–5 years of age chronically malnourished (Giese et al., 2022).

2.2. Childcare facilities and child development data

In 2021, the Thrive by Five Index study was conducted to generate a national estimate for the Early Learning Outcomes Measure (ELOM) (Giese et al., 2022). In the study, 1247 childcare facilities were surveyed using a three-stage cluster sampling approach. Clusters of childcare facilities were randomly selected, followed by a random selection of childcare facilities within each cluster, and finally, a random selection of 3–5 children, aged 50–59 months, per childcare facility. From this dataset, a national estimate of 65% of children are not 'on track' for physical and cognitive development (defined as meeting learning standards, conduct tasks expected for their age, and meet growth standards). In addition, a robust dataset, including GPS locations and child development measures of each childcare facility, was generated. In total, 432 clusters, 1247 childcare facilities, and 5184 children were surveyed. Additional details on the national study can be found in Giese et al., (2022) (Giese et al., 2022).

2.3. GSV pilot study

For this pilot study, we randomly selected 43 childcare facilities from the top and bottom 10% of the original 432 clusters based on mean overall child development scores for the cluster (N = 86 total childcare facilities). This initial sample size was selected based on necessary sample sizes to detect significant differences in the alpha and beta scores, previously demonstrated sample sizes used in the LEfSE analysis (Segata et al., 2011a), as well as the cost of the data and analytic tools. Facilities were labeled as 'high' or 'low' based on if they were selected from the top 10% or bottom 10% of child development scoring clusters, respectively. Next, each site which hosted a childcare facility was evaluated for LEHC label abundances. Here, abundance refers to the number of times a specific label is found within a site.

2.4. GSV analysis and abundance estimates

To create a standard approach for generating LEHC label abundance estimates for each site that hosted a childcare facility, a 1600-m square geofence was set with the GPS coordinates for the childcare facility in the center. This distance was selected based on the range of previously reported data evaluating associations between pollution (noise, air, water, soil, etc.) and child health and the significantly diminished effects as the distance between the children and the point and non-point sources increases (Maantay and McLafferty, 2011). In addition, 1600m provides a robust set of images in which to characterize the neighborhood environment. Next, based on the field of view for each Google's Street View (GSV) image, GPS coordinates were then generated every 30 m apart covering the entire geofence site. GPS coordinates less than 30 m apart may generate the same image due to the spacing of images within the GSV database, increasing the risk of double counting and amplifying representation biases. The set of GPS coordinates was then uploaded to the Google Cloud Shell Platform to temporarily download the images for analysis using the GSV Application Programming Interface (API). Four images per GPS coordinate were downloaded, each with a unique, equally spaced heading (0º, 90˚, 180˚, 270˚) but having the same pitch (5% decline). The GSV API uses the GPS coordinates, heading, and pitch to check if a GSV image is available for that specific location. Availability is based on if a road is present and if an imaging vehicle or person has submitted an image from that location. If the image was available, it was automatically downloaded into a folder specific to that childcare facility. Due to a higher abundance of available GSV images for high versus low scoring sites, GSV images for high scoring sites were randomly subsampled at 23% of the total images available. Images from N = 86 geofenced sites with a childcare facility were downloaded. However, in both groups, if the total number of images for a site was below 240, it was removed from the analysis. Based on preliminary evaluations and the mean number of images per site, sites under this cutoff may not be sufficiently represented. This cutoff value produced N = 60 sites (n = 30 in each group) for further processing.

Next, using the Google Cloud Shell Platform, the Google Vision AI API was called to analyze each image using the label detection models. The label detection models are proprietary, but often are a set of deep convolution neural networks (Nguyen et al., 2019). For Google Vision AI, they can generate >10,000 unique image labels and provide up to 50 labels per image along with an accuracy confidence score for each. The accuracy confidence score threshold for the labels was set at 60%. Previous validation studies have demonstrated high concordance between the default cutoff (50%) and hand-labeled images for a subset of labels (Nguyen et al., 2019). Given our full-data approach, we set the cutoff at 60%. Example labels include items such as car, bicycle, or wheel for transportation or shrubland, prairie, or sand for landscape (see Fig. 1 for an example). The labels can vary widely, and, similarly to bioinformatics data, form part of a hierarchy or taxonomy. However, this taxonomy does not yet exist for the set of algorithms in the Google Vision AI API label detection tool. Therefore, we followed Hlava (2015) in developing an initial version of a Google Vision AI label taxonomy (Hlava, 2015). This included the following steps: 1) setting boundary conditions, 2) defining the core components of the taxonomy (the highest level of taxa), 3) taxonomy creation, and 4) taxonomy validation. The 1100 most prevalent individual labels from across all geofenced sites were retained and a taxonomy was created. To create the taxonomy, two researchers were given the list of labels, a small example set, and a list of labels designated by the authors as the highest-level taxonomy (i.e., core compo-

Label	Accuracy
Sky	97%
Plant	92%
Ecoregion	92%
Cloud	91%
Land Lot	86%
Asphalt	86%
Road Surface	84%
Grass	79%
Landscape	78%
Road	75%

Label	Accuracy
Auto Parking Light	98%
Car	97%
Plant	96%
Wheel	96%
Auto Side Marker Light	96%
Tire	95%
Vehicle	95%
Sky	93%
Tree	89%
Window	87%

Fig. 1. Example of output of Google Vision AI analysis.
Top 10 labels listed based on accuracy scores (study cutoff was 60%). Total labels with ≥60% accuracy generated for the left image was n = 20 while the right image was n = 40. Images by Lee E. Voth-Gaeddert.

nents). The two researchers independently created a taxonomy, the two taxonomies were then compared, and any discrepancy was resolved by the first author of this paper (see the supplementary material for further details). We provide additional definitions and terms common within the field of molecular bioinformatics, for those readers less familiar (see Text S1).

Image labels were aggregated for each geofenced site and a data matrix of raw counts for each label and site was created. To normalize the data, established bioinformatic methods from Bolyen et al. were adapted and followed (Bolyen et al., 2019). Briefly, the total raw count of labels for each site was divided by the median total number of labels across all sites resulting in a normalized raw count for all geofenced sites. Finally, to test if changing the initial heading altered the abundance estimates, a sub-analysis was conducted to compare the label count estimates for a small set of GPS coordinates (n = 96) at 0, 90, 180, and 270 to the same set of GPS coordinates but with a heading of 45, 135, 225, and 315 (see supplementary material Text S2 and Fig. S1 for further details).

2.5. Data analysis

The normalized raw counts of labels for each area were then used to evaluate the differences in childcare facilities that had a child development score in the top 10% ('high') to those in the bottom 10% ('low'). First, to visually evaluate abundance distributions across sites and groups, a bar plot was generated. Due to the high complexity of the label abundances, only the highest taxonomic level was depicted (Level 1). The bar plot was estimated using the plot_bar function in the Phyloseq package in R (default parameters (Mcmurdie and Holmes, 2013);).

Second, alpha and beta diversity differences were evaluated. For alpha diversity, the Shannon index was used to evaluate differences visually via a box-and-whisker-plot and statistically via the pairwise Wilcox Test. For beta diversity, due to the presence of null values in the dataset (some sites having zero of a specific label), the Bray-Curtis distance metric was used to evaluate dissimilarities between groups visually by non-metric Multidimensional Scaling (NMDS) plots and statistically via PERMANOVA by adonis (Oksanen, 2022). In addition, ellipses denoting the two-dimensional t-distribution and normal distribution (dotted line), were added to the NMDS plot to visually evaluate differences in high and low groups. This visual differentiation was also used to identify a subset of sites for further analyses (see Table S1). This included a positive deviance analysis ('high' scoring sites in the 'low' scoring region of the NMDS plot), negative deviance analysis ('low' scoring sites in the 'high' scoring region of the NMDS plot), and an inter-comparison analysis (a set of 'high' and 'low' scoring sites that occupied the same region of the NMDS plot). All analyses and plots described above were conducted in R using the Phyloseq package (Mcmurdie and Holmes, 2013).

To evaluate label-specific differences between high and low scoring sites, Linear discriminant analysis Effect Size (LEfSe) was used. The LefSe function on Galaxy Huttenhower platform was used with default parameters (Segata et al., 2011b). An analysis of all neighborhoods, grouped in high vs low scoring groups was conducted and individual feature analysis, bar plots, and cladograms were generated. Next, three sub-analyses were conducted including positive and negative deviance, and an inter-comparison. Further details on the specific sites used for the deviance analyses and inter-comparison are presented in Table S1 in the supplementary material.

3. Results

Twenty-six of the selected 86 sites had fewer than 240 images within the geofence and were dropped from the analysis. Normalized label count estimates for N = 60 geofenced sites with childcare facilities were generated, n = 30 from the top and bottom 10% of sampling clusters for child development scores, respectively. Table 1 presents a set of descriptive statistics for the high and low scoring sites. While the basic descriptive statistics were similar between groups, the composite child development scores were intentionally different between groups. In addition, the available images per site and generated labels per site were different, with twice as many images and labels available from high scoring sites compared to low scoring sites. Hence a normalization step was critical.

The bar plot in Fig. 2, depicts the abundance of labels at the first or highest level of the taxonomy for each individual site, grouped by high and low scores. Labels associated with infrastructure and landscape were the most prevalent for all sites. Labels associated with the city, infrastructure, and transport appear to have a higher abundance among high scoring sites, where labels associated with animals, landscape, and sky appear to have a higher abundance among low scoring sites. The LefSe analysis presented below provides details on label-specific differences at lower taxonomic levels.

To compare overall differences in labels between groups, alpha and beta diversity metrics were estimated. Fig. 3a depicts the box-and-whisker plot for the alpha diversity measure (Shannon Index). The alpha diversity measure was statistically different between high and low scoring sites according to the pairwise Wilcox Test (p-value <0.001). This suggests that there was a higher diversity of labels among high scoring sites.

Fig. 3b is the ordination plot which visually depicts the dissimilarities in label counts between grouped sites based on the Bray-Curtis dissimilarity metric. The two ellipses denote the t-distribution (solid line) and normal distribution (dotted line) for each group of sites. The figure suggests that based on labels, high scoring sites form a specific cluster which partially overlaps with the low scoring sites. In addition, there appears to be a specific point horizontally (0.1 on the x-axis) where the high and low scoring

Table 1
Descriptive Statistics for the childcare facilities in the final set of geofenced neighborhoods.

Description	High Scoring (Top 10%) Mean (Std)	Low Scoring (Bottom 10%) Mean (Std)
Altitude (meters)	904 (688)	1043 (457)
Sex (prevalence)	53% girls	54% girls
Age (months)	55.1 (1.65)	54.2 (1.51)
Height (cm)	106.6 (4.11)	103.5 (2.52)
Height-for-age z-score	−0.10 (0.82)	−0.67 (0.52)
Cluster Composite ECD Score	59.0 (4.81)	32.0 (3.71)
Facility Composite ECD Score	60.7 (9.47)	30.3 (8.05)
Domain 1 (GMD)	9.72 (2.04)	5.73 (2.02)
Domain 2 (FMCVMI)	13.9 (2.15)	8.30 (1.98)
Domain 3 (ENM)	11.5 (2.51)	5.37 (2.22)
Domain 4 (CEF)	11.5 (3.28)	3.96 (1.60)
Domain 5 (ELL)	14.1 (2.14)	6.91 (3.24)
Images per neighborhood (Median, range)	1655 (329–2237)	850 (248–3756)
Labels per neighborhood (Median, range)	48,217 (9442–69509)	22,870 (7198–93692)

GMD = gross motor development; FMCVMI = fine motor coordination and visual motor integration; ENM = emergent numeracy and mathematics; CEF = cognition and executive functioning; ELL = emergent literacy and language.

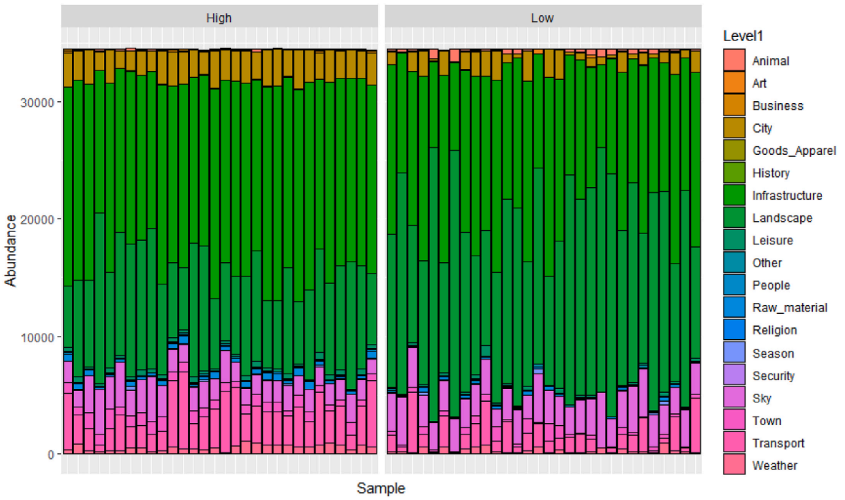

Fig. 2. Label abundance bar plot of individual sites grouped by high and low scores.

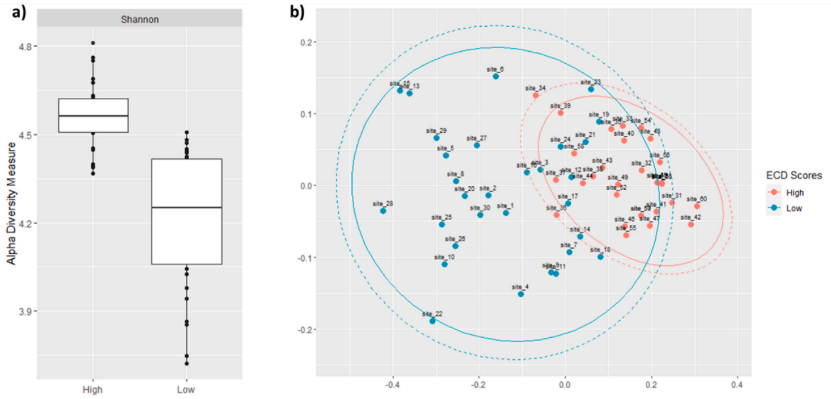

Fig. 3. a) Box plot of alpha diversity of labels from neighborhoods and b) ordination plot of neighborhoods.
High and low scores refer to ECD scores of childcare facilities in a site. Shannon index used for alpha diversity. Ordination plot depicting beta diversity is a non-metric Multidimensional Scaling (NMDS) plot using the Bray-Curtis dissimilarity metric.

sites differentiate themselves. Statistically, there is a significant difference between the two groups, based on the PERMANOVA test ($p < 0.001$). Fig. 3b was also used to identify positive and negative deviant sites as well as a set of sites for inter-comparison between the two groups. The list of neighborhoods included in all three sub-analyses can be found in the supplementary material.

LefSe analysis was used to evaluate significantly different labels between 1) overall high and low scoring sites, 2) positive deviant sites, 3) negative deviant sites, and 4) inter-comparison sites. Fig. 4, Fig. S2, and Table 2 report the key significant differences across the four comparisons. The overall comparison had 388 individual labels that were significantly different. Fig. 3 aggregates these at levels 1 and 2, while Table 2 describes the categories of the top 30 most significantly different labels. For example, Table 2 reports that labels associated with 'transport' were much more abundant among high scoring sites. Fig. 3 suggests this was primarily driven

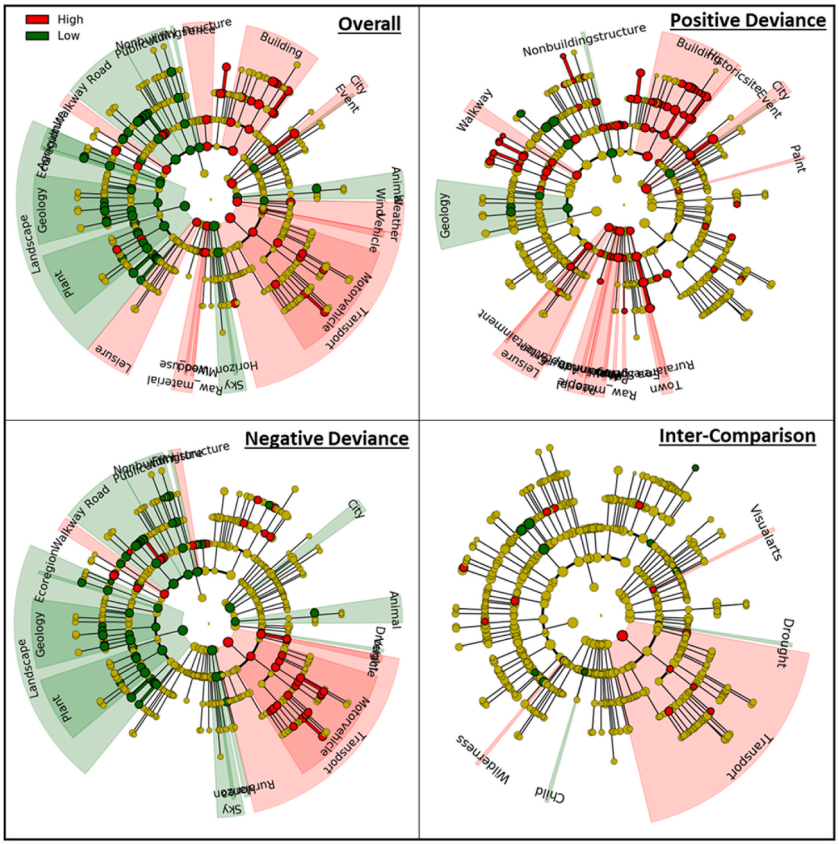

Fig. 4. LefSe Cladogram for each comparison.
There were seven taxonomic levels, names and color wedges only shown for levels 1 and 2 for clarity (further details in the supplementary material). Sample sizes include overall (30 high vs 30 low), positive deviance (5 high vs 23 low), negative deviance (25 high vs 7 low), and inter-comparison (9 high vs 9 low).

Table 2
Summary of Top 30 Abundant Labels across all comparisons.

	High	Low
Overall	Transport *Cars, vans, trucks, buses, bicycles, etc.* Residential Buildings *Homes, apartments, town houses, condominiums, etc.*	Roads *Dirt roads, asphalt, tar, road surface, highway* Dry and Rural Land *Field, soil, sand, grass, plain, prairie, shrubland* Electrical public utilities *Electrical network, overhead powerlines*
Positive Deviance	Residential Buildings and Neighborhoods *Homes, cottage, apartment, residential area, real estate, etc.*	Roads *Asphalt, tar, road surface* Dry and Rural Land *Plateau, badlands, geology*
Negative Deviance	Transport *Cars, vans, trucks, buses, bicycles, etc.*	Roads *Dirt roads, asphalt, tar, road surface, highway* Dry and Rural Land *Field, soil, sand, grass, plain, prairie, shrubland, aeolian dunes* Electrical public utilities *Electrical network, overhead powerlines, electricity*
Inter-Comparison	Transport *Cars, vans, trucks, buses, bicycles, etc.* Moist Landscape *Wetlands, water, lacustrine plain*	Roads *Asphalt, tar, road surface* Dry and Rural Land[a] *Shrubland, drought, canopy* Commercial and Industrial Buildings *Commercial building, garage door, building material*

[a] Labels associated are slightly different compared to other categories with the same name. Total number of statistically different labels varied by comparison, overall: 388; positive deviant: 123; negative deviant: 212; inter-comparison: 25.

by labels associated with motor vehicles (cars, buses, motorcycles, etc.). Additionally, 'buildings' and specifically those classified as residential buildings were more abundant among high scoring sites. For low scoring sites, 'landscape' labels associated with dry and rural characteristics were much more abundant along with roads and large electrical public utility labels.

The positive deviance analysis compared five high scoring and 23 low scoring sites, the negative deviance analysis compared 25 high scoring and seven low scoring sites, and the inter-comparison compared 18 evenly divided and grouped sites between high and low scores. The Cladograms in Fig. 4, as well as the category labels summarized in Table 2, allowed for the identification of unique dynamics. For example, the most important attribute that differentiated the positive deviant sites (i.e., high scoring) was residential buildings, as also seen in the overall analysis. This may suggest that if a childcare facility is embedded in a community, as opposed to a commercial or industrial location, better child development outcomes are possible. This is further supported by the higher abundance of commercial and industrial labels among low scoring sites in the inter-comparison. However, the residential building attribute disappears when evaluating what was missing among low scoring sites in the negative deviance analysis. The key attribute that was missing in low scoring sites (i.e., labels in high abundance in high scoring sites) was transportation. This attribute was also a key differentiator when conducting the inter-comparison. This may suggest that the availability of transport near childcare facilities is an important attribute for improved child development outcomes.

Labels in higher abundance among low scoring sites were also relatively consistent across comparisons. All four comparisons identified labels associated with 'roads' and 'dry and rural land.' Road labels included characteristics such as road surface, dirt roads, asphalt, tar, highways, etc. While all GSV images for all sites were taken from the road, low scoring sites had a higher abundance of labels associated with roads. This may suggest that there were either fewer structures (buildings or vehicles) to impede seeing other roads, fewer other labels to attribute to an image, or longer stretches of road in the same direction. Dry and rural land labels included characteristics such as grasslands, prairies, shrubland, sand, and soil, among others. This suggests that childcare facilities were near open areas and less likely to be embedded in communities. However, this may be due to the low population density or overall total number of households in an area. Finally, in the overall comparison and the negative deviance analysis, labels affiliated with electrical public utilities were more abundant. This may support the previous reported findings that embedding childcare facilities in residential or community-centered areas, as opposed to commercial or industrial areas, is important for child development outcomes.

4. Discussion

Machine learning and big data can play an important role in improving our ability to identify and mitigate potential environmental exposures to communities at a global scale while maintaining local granularity. In this study, we demonstrate a unique, big data approach to exploring potential environmental differences among sites that host childcare facilities and the corresponding child development outcomes. This approach provided an exploratory method to identify potential local environmental health characteristics associated with differences in child development outcomes in childcare facilities. While previous applications have demonstrated the utility of GSV images and computer vision models, our approach provides a fundamental reframing of how we apply analytical methods to represent high dimensional or "big" data. Here we maximized the number of input labels to provide a 'full data' approach to improve our understanding of the association between environmental factors and child development outcomes. Each bioinformatics tool applied in this study provided a unique way to analyze and visualize different characteristics of the high dimensional data and

the associations with child development outcomes such as diversity metrics, distance metrics, or effect sizes. Previous literature has demonstrated the complexity of environmental factors associated with improved child development outcomes. Therefore, having a method that can help reduce this initial complexity (i.e., data reduction approaches), can be a powerful tool to incorporate into a larger series of steps for understanding factors driving child development outcomes.

Using this approach, we find that a higher abundance of image labels associated with transportation and residential or community spaces are present in high scoring sites. In low scoring sites, labels associated with roads, dry and rural land, and electrical public utilities were significantly more abundant. Interestingly, for positive deviant sites (those with high scores but that clustered with low scoring sites), the data suggested that there was a significantly higher abundance of residential buildings near the childcare facilities with higher child development scores. This was also supported by the higher abundance of commercial and industrial labels among low scoring sites in the inter-comparison analysis. For negative deviant sites (those with low scores but that clustered with high scoring sites), the data suggested that there was a significantly higher abundance of labels associated with transportation in the high scoring sites. This suggested that the lack of transportation labels (primarily motor vehicles such as cars) among low scoring sites may be associated with lower child development scores. This was also supported by the results from the inter-comparison analysis.

While child development metrics have been demonstrated to be associated with the household and childcare facility environments, the community or neighborhood environments are also hypothesized to be associated with child development (Anders et al., 2012; Black et al., 2017; Walker et al., 2011). Walker and colleagues, in a Lancet series, summarize substantial work around elucidating the drivers to improved child development outcomes, suggesting there are many direct (e.g., malnutrition and environmental exposures) and indirect effects (e.g., poverty) that can lead to poor outcomes (Walker et al., 2011). Furthermore, as demonstrated by Anders and colleagues, these direct and indirect effects can vary over time in significance and magnitude (Anders et al., 2012). Ecological systems theory has been used to help understand the role of various social and physical systems in which early childcare is embedded and the effect of the various attributes of these systems on child development outcomes (Marshall, 2004). Our findings align with previous work demonstrating the importance of the larger built environment surrounding the childcare facility, including hosting the childcare facility in more residential areas (Loeb et al., 2004) and improving transportation access in those areas (Satoh et al., 2018; Walker et al., 2011).

The proposed method in this study is exploratory and should be paired with locally targeted confirmatory approaches to limit overinterpretation of results. For example, within the South African context, a potential next step could include reevaluating the 60 sites, using a more targeted approach to quantifying the percent area surrounding the childcare facility classified as residential. This could be integrated into a larger regression model to better understand the contribution of location in explaining the variance of the child development outcome. Previous research in South Africa has suggested that physical activity is associated with positive child development outcomes and that the local environment can play an important role in promoting such activity (Uys et al., 2016). For transportation, further evaluation could be conducted by looking at what aspects of transportation have positive impacts on child development (is it public transport like buses, or walkways, that drive improved outcomes?). In South Africa, seminal work has been conducted looking at public transportation, income poverty, and children (Lucas, 2011). Furthermore, this method may be able to predict child development outcomes, given a set of new study sites or neighborhoods as it provides a method that is not a complete black box. Finally, child development outcomes are aggregated based on a set of sub-domains each with their own unique areas. Further analyses in domains related to physical development may yield information helpful to understanding the role of the environment on child health. Regardless, this method should be applied appropriately, as an exploratory tool, and complemented with confirmatory approaches.

There are a number of limitations with this type of approach, many of which have been discussed in detail in the context of molecular bioinformatics (Escobar-Zepeda et al., 2015; Hiraoka et al., 2016) and previous GSV applications (Rzotkiewicz et al., 2018). First, the GSV images available in a given site can vary widely and are often correlated with socio-economic status (as this is associated with general infrastructure development). To account for this, we have used well established normalization techniques, however, diversity metrics sensitive to total unique labels will have a bias towards high scoring sites. Second, we used computer vision algorithms trained and provided by the Google Vision AI team via their API. While generally it is known that convoluted neural networks are used, various details on training quality, hyperparameters, and similar metrics are unknown. In addition, certain types of algorithms may be more prevalent (i.e., there may be more transportation related algorithms versus art algorithms), which increases the risk of overrepresentation in the taxonomy. Third, the image labels were not mutually exclusive to each other and therefore a taxonomy was required to provide a hierarchical structure to the labels. This taxonomy was built by the researchers and not the team which runs Google Vision AI. Therefore, misinterpretation of some labels is possible, however, most results reported here were significantly different at a high taxonomic level (level 1 or 2). Fourth, spatial autocorrelation can introduce bias into results, however, the three-stage sampling technique used in the original childcare survey minimizes this risk. Finally, we used the top and bottom 10% of clusters of childcare facilities to demonstrate our technique, however, this can hide important nuances within the remaining clusters and this population segment should be evaluated in future research. Despite these challenges, we aim to demonstrate the utility of this approach as one tool in a set of tools that can be used to better understand the role of the environment on various outcomes, such as child development.

Finally, child development is one of many outcome metrics that can be evaluated. In the introduction section, we highlighted other outcomes such as injury, violence, alcohol and tobacco use, physical activity, mental health (Rzotkiewicz et al., 2018), COVID-19 cases (Nguyen et al., 2020), mental health metrics (Odgers et al., 2012), acute and chronic disease outcomes, and mortality. These metrics could replace our current child development metric to allow for a granular evaluation of drivers to those societal issues. In addition, metrics such as disaster preparedness or climate change resilience could be evaluated to better understand how communities can prepare for future challenges.

5. Conclusion

In this study, we integrated GSV images and computer vision models with bioinformatic approaches to evaluate significant differences in LEHC labels of the built and natural environment between high and low child development scoring sites hosting childcare facilities. The approach provided a big data method that could account for the high level of complexity from the environment. The results suggested that higher abundances of transportation and residential buildings were associated with higher child development scores, while more rural, dry, and industrial attributes were associated with lower child development scores. Combined with previous literature, the data suggest that geographic location and mobility are important decision criteria for new or existing childcare facilities. To continue to improve the method, the taxonomy should be refined (with possible input from the team at Google), methods to reduce bias in GSV image representation should be further improved, and additional output metrics should be explored. However, this method can provide a granular approach to understanding unique local dynamics while being scaled globally.

Funding sources

This work was supported by Innovation Edge.

Author contribution statement

LVG: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft. XH: Conceptualization; Investigation; Project administration; Resources; Writing – review & editing. MT: Conceptualization; Funding acquisition; Methodology; Resources; Supervision; Writing – review & editing. AVH: Conceptualization; Funding acquisition; Methodology; Resources; Supervision; Writing – review & editing.

Ethical statement

All ethical practices have been followed in relation to the development, writing, and publication of the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is accessible via the Google Street View API.

Acknowledgements

The authors are grateful for the support provided by Steyn Lodewyk Vogel.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rsase.2023.100950.

References

Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J.J., Tripathi, A., Brenner, D.A., Loomba, R., Smarr, L., Sandborn, W.J., Schnabl, B., Dorrestein, P., Zarrinpar, A., Knight, R., 2019. Microbiome 101: studying, analyzing, and interpreting gut microbiome data for clinicians. Clin. Gastroenterol. Hepatol. 17, 218–230. https://doi.org/10.1016/J.CGH.2018.09.017.

Anders, Y., Rossbach, H.G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., von Maurice, J., 2012. Home and preschool learning environments and their relations to the development of early numeracy skills. Early Child. Res. Q. 27, 231–244. https://doi.org/10.1016/J.ECRESQ.2011.08.003.

Andres, L., Boateng, K., Borja-Vega, C., Thomas, E., 2018. A review of in-situ and remote sensing technologies to monitor water and sanitation interventions. 2018. Water 10. https://doi.org/10.3390/W10060756. Page 756 10, 756.

Barbarin, O.A., Richter, L.M., 2001. Mandela's Children: Growing up in Post-apartheid South Africa, Mandela's Children: Growing up in Post-apartheid South Africa. Routledge, New York, NY, US.

Black, M.M., Walker, S.P., Fernald, L.C.H., Andersen, C.T., DiGirolamo, A.M., Lu, C., McCoy, D.C., Fink, G., Shawar, Y.R., Shiffman, J., Devercelli, A.E., Wodon, Q.T., Vargas-Barón, E., Grantham-McGregor, S., 2017. Early childhood development coming of age: science through the life course. Lancet 389, 77–90. https://doi.org/10.1016/S0140-6736(16)31389-7.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K. Bin, Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

Britto, P., 2015. Why early childhood development is the foundation for sustainable development [WWW Document]. UNICEF Connect. URL. https://blogs.unicef.org/blog/why-early-childhood-development-is-the-foundation-for-sustainable-development/.

Capone, D., Adriano, Z., Berendes, D., Cumming, O., Dreibelbis, R., Holcomb, D.A., Knee, J., Ross, I., Brown, J., 2019. A localized sanitation status index as a proxy for fecal contamination in urban Maputo, Mozambique. PLoS One 14, e0224333. https://doi.org/10.1371/JOURNAL.PONE.0224333.

Dave, M., Higgins, P.D., Middha, S., Rioux, K.P., 2012. The human gut microbiome: current knowledge, challenges, and future directions. Transl. Res. 160, 246–257. https://doi.org/10.1016/J.TRSL.2012.05.003.

Escobar-Zepeda, A., De Leon, A.V.P., Sanchez-Flores, A., 2015. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front. Genet. 6, 1–15. https://doi.org/10.3389/fgene.2015.00348.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Fei-Fei, L., 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. Proc. Natl. Acad. Sci. U. S. A 114, 13108–13113. https://doi.org/10.1073/PNAS.1700035114/SUPPL_FILE/PNAS.1700035114.SAPP.PDF.

Giese, S., Dawes, A., Tredoux, C., Mattes, F., Bridgman, G., van der Berg, S., Schenk, J., Kotze, J., 2022. Thrive by Five Index Report. Cape Town.

Hiraoka, S., Yang, C.-C., Iwasaki, W., 2016. Metagenomics and bioinformatics in microbial ecology: current status and beyond. Microb. Environ. 31, 204–212. https://doi.org/10.1264/jsme2.ME16024.

Hlava, M.M.K., 2015. The taxobook: principles and practices of building taxonomies. In: The Taxobook. Morgan & Claypool. first ed., Chapel Hill. https://doi.org/10.1007/978-3-031-02288-3.

Jeon, D.J., Ligaray, M., Kim, M., Kim, G., Lee, G., Pachepsky, Y.A., Cha, D.H., Cho, K.H., 2019. Evaluating the influence of climate change on the fate and transport of fecal coliform bacteria using the modified SWAT model. Sci. Total Environ. 658, 753–762. https://doi.org/10.1016/J.SCITOTENV.2018.12.213.

Jones, K.H., Daniels, H., Heys, S., Ford, D.V., 2018. Challenges and Potential Opportunities of Mobile Phone Call Detail Records in Health Research: Review, 6. JMIR Mhealth Uhealth. https://doi.org/10.2196/MHEALTH.9974.

Koletzko, B., Godfrey, K.M., Poston, L., Szajewska, H., Van Goudoever, J.B., De Waard, M., Brands, B., Grivell, R.M., Deussen, A.R., Dodd, J.M., Patro-Golab, B., Zalewski, B.M., 2019. Nutrition during pregnancy, lactation and early childhood and its implications for maternal and long-term child health: the early nutrition Project recommendations. Ann. Nutr. Metab. 74, 93–106. https://doi.org/10.1159/000496471.

Loeb, S., Fuller, B., Lynn Kagan, S., Carrol, B., 2004. Child care in poor communities: early learning effects of type, quality, and stability. Child Dev. 75, 47–65. https://doi.org/10.1111/J.1467-8624.2004.00653.X.

Lu, Y., Treiman, D.J., 2011. Migration, remittances and educational stratification among blacks in apartheid and post-apartheid South Africa. Soc. Forces 89, 1119–1143. https://doi.org/10.1093/sf/89.4.1119.

Lu, Y., 2019. Using Google Street View to investigate the association between street greenery and physical activity. Landsc. Urban Plann. 191, 103435. https://doi.org/10.1016/J.LANDURBPLAN.2018.08.029.

Lucas, K., 2011. Making the connections between transport disadvantage and the social exclusion of low income populations in the Tshwane Region of South Africa. J. Transport Geogr. 19 (6), 1320–1334. https://doi.org/10.1016/j.jtrangeo.2011.02.007.

Maantay, J.A., McLafferty, S., 2011. Environmental health and geospatial analysis: an overview. In: Geospatial Analysis of Environmental Health. Springer, Dordrecht, pp. 3–37. https://doi.org/10.1007/978-94-007-0329-2_1.

Marshall, N.L., 2004. The quality of early child care and children's development. Curr. Dir. Psychol. Sci. 13, 165–168. https://doi.org/10.1111/J.0963-7214.2004.00299.X.

Mcmurdie, P.J., Holmes, S., 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8. https://doi.org/10.1371/journal.pone.0061217.

Nguyen, Q.C., Huang, Y., Kumar, A., Duan, H., Keralis, J.M., Dwivedi, P., Meng, H.W., Brunisholz, K.D., Jay, J., Javanmardi, M., Tasdizen, T., 2020. Using 164 million google street view images to derive built environment predictors of COVID-19 cases. Int. J. Environ. Res. Publ. Health 17, 1–13. https://doi.org/10.3390/IJERPH17176359.

Nguyen, Q.C., Khanna, S., Dwivedi, P., Huang, D., Huang, Y., Tasdizen, T., Brunisholz, K.D., Li, F., Gorman, W., Nguyen, T.T., Jiang, C., 2019. Using Google Street View to examine associations between built environment characteristics and U.S. health outcomes. Prev. Med. Reports 14, 100859. https://doi.org/10.1016/J.PMEDR.2019.100859.

Odgers, C.L., Caspi, A., Bates, C.J., Sampson, R.J., Moffitt, T.E., 2012. Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method. JCPP (J. Child Psychol. Psychiatry) 53, 1009–1017. https://doi.org/10.1111/J.1469-7610.2012.02565.X.

Oksanen, J., 2022. Vegan: Community Ecology Package.

Pickering, A.J., Julian, T.R., Marks, S.J., Mattioli, M.C., Boehm, A.B., Schwab, K.J., Davis, J., 2012. Fecal contamination and diarrheal pathogens on surfaces and in soils among Tanzanian households with and without improved sanitation. Environ. Sci. Technol. 46, 5736–5743. https://doi.org/10.1021/ES300022C.

Rundle, A.G., Bader, M.D.M., Richards, C.A., Neckerman, K.M., Teitler, J.O., 2011. Using google street view to audit neighborhood environments. Am. J. Prev. Med. 40, 94. https://doi.org/10.1016/J.AMEPRE.2010.09.034.

Russell, A.L., Hentschel, E., Fulcher, I., Ravà, M.S., Abdulkarim, G., Abdalla, O., Said, S., Khamis, H., Hedt-Gauthier, B., Wilson, K., 2022. Caregiver parenting practices, dietary diversity knowledge, and association with early childhood development outcomes among children aged 18-29 months in Zanzibar, Tanzania: a cross-sectional survey. BMC Publ. Health 22, 1–14. https://doi.org/10.1186/S12889-022-13009-Y/TABLES/6.

Rzotkiewicz, A., Pearson, A.L., Dougherty, B.V., Shortridge, A., Wilson, N., 2018. Systematic review of the use of Google Street View in health research: major themes, strengths, weaknesses and possibilities for future research. Health Place 52, 240–246. https://doi.org/10.1016/J.HEALTHPLACE.2018.07.001.

Satoh, K., Tsukahara, K., Yamamoto, K., Satoh, K., Tsukahara, K., Yamamoto, K., 2018. Location evaluation of childcare facilities focusing on transportation in Japanese urban areas. J. Geogr. Inf. Syst. 10, 521–538. https://doi.org/10.4236/JGIS.2018.105028.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011a. Metagenomic biomarker discovery and explanation. Genome Biol. 12, 1–18. https://doi.org/10.1186/gb-2011-12-6-r60.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011b. Metagenomic biomarker discovery and explanation. Genome Biol. 12, 1–18. https://doi.org/10.1186/gb-2011-12-6-r60.

Statistics South Africa, 2020. Child Poverty in South Africa: A Multiple Overlapping Deprivation Analysis. Pretoria.

Uys, M., Bassett, S., Draper, C.E., et al., 2016. Results from South Africa's 2016 report card on physical activity for children and youth. J. Phys. Activ. Health 13 (11 Suppl. 2), S265–S273. https://doi.org/10.1123/jpah.2016-0409.

Voth-Gaeddert, L.E., Cudney, E.A., Oerther, D.B., 2018a. Primary factors statistically associated with diarrheal occurrences. Environ. Eng. Sci. 35, 836–845. https://doi.org/10.1089/ees.2017.0338.

Voth-Gaeddert, L.E., Stoker, M., Cornell, D., Oerther, D.B., 2018b. What causes childhood stunting among children of San Vicente, Guatemala: employing complimentary, system-analysis approaches. Int. J. Hyg Environ. Health 221, 391–399. https://doi.org/10.1016/j.ijheh.2018.01.001.

Voth-Gaeddert, L.E., Torres, O., Maldonado, J., Krajmalnik-Brown, R., Rittmann, B.E., Oerther, D.B., 2019. Aflatoxin exposure, child stunting, and dysbiosis in the intestinal microbiome among children in Guatemala. Environ. Eng. Sci. 36. https://doi.org/10.1089/ees.2019.0104.

Walker, S.P., Wachs, T.D., Grantham-Mcgregor, S., Black, M.M., Nelson, C.A., Huffman, S.L., Baker-Henningham, H., Chang, S.M., Hamadani, J.D., Lozoff, B., Gardner, J.M.M., Powell, C.A., Rahman, A., Richter, L., 2011. Inequality in early childhood: risk and protective factors for early child development. Lancet 378, 1325–1338. https://doi.org/10.1016/S0140-6736(11)60555-2.