Applications of Geospatial Data and Methods in Environmental Epidemiology

S.M. Labib
Department of Geography, School of Environment, Education and Development
University of Manchester.

smlabib.com

20-02-2020
Who Am I?

I am Labib.

From Bangladesh, travelled from Manchester.

My Academic Journey:

- BSc in Urban and Regional Planning (2014), Bangladesh University of Engineering and Technology.
- MSc in Geographical Information Science (2017), University of Manchester.
- PhD in Physical Geography (2017-2020, submission), University of Manchester.

Research Interests:

Geographic Information Science, Remote Sensing, Green Infrastructure, Transportation planning, and Environmental Epidemiology.
Content & Outline

• **Brief Overview of geospatial data and methods in epidemiology (5 min)**
  - Historic example and now
  - Geospatial approaches in practice
  - Exposure assessment

• **Geospatial Data and methods case studies (10 min)**
  - Airborne imagery data,
  - OpenStreeMap data,
  - Volunteer GIS data

• **Examples of model coupling and their applications (10 min)**
  - Applying in transportation sustainability
  - Combining machine learning models with spatial data

• **Spatial dimensions in greenspace and health research- a systematic review (20 min)**
  - Scale
  - Exposure assessment (data, methods)
  - The buffering issue
  - MAUP and spatial autocorrelation

• **Q-A? (15 min)**

• **References**
Part -1: Brief Overview of geospatial data and methods in epidemiology
1854 Broad Street cholera outbreak

- What is the **average distance from the contaminated pump** to the surrounding locations?
- What is there now? [http://tiny.cc/9j17jz](http://tiny.cc/9j17jz)


**Full Story:** [https://youtu.be/lNjrAXGRda4](https://youtu.be/lNjrAXGRda4)
Geospatial approaches in practice

- **Environmental Factors**: Pollution sources (e.g., air, water pollution), natural environment, built environment. **Spatial Data dominance!**
- **External influence measurement**: Exposure assessment - a function of location (proximity) and time (Nieuwenhuijsen, 2009). **Spatial Methods dominance!**
Part -2: Geospatial Data (case studies)
Geospatial Data (case studies)

Case study 1: Satellite imagery data

The potentials of Sentinel-2 and LandSat-8 data in green infrastructure extraction, using object based image analysis (OBIA) method

S M Labib & Angela Harris

- Low availability of greenspace data in Dhaka, the existing data are usually outdated.
- New free satellite data from improved sensors are available (Sentinel-2, 10m), Landsat-8 (30m)
- Which performs better in extracting greenspace better, what are the issues?
• Applied **Object based image analysis**. A semi-automation process
• **Sentinel-2 had greater accuracy** (71.24%) in detecting greenspace, buildings; than Landsat-8 (67.85%)
Case study 2: OpenStreetMap data (a pilot test)

- Largest open access crowdsourced Geo-data
- Global coverage of street network, integrated in Global Roads Inventory Project (GRIP) dataset.
- Has anonymized GPS tracks up to 2013, global coverage (>21 GB of GPS points)
- Can such GPS data be useful for understanding urban Park usage?

Source: Meijer et al., (2018)

Data Source: https://www.globio.info/download-grip-dataset
Geospatial Data (case studies)

Case study 2: OpenStreetMap data (a pilot test)

Data Source: https://planet.openstreetmap.org/gps/
Geospatial Data (case studies)

Case study 2: OpenStreetMap data (a pilot test)

Connecting GPS points using GRASS GIS

Connecting GPS points using SAGA GIS

Connecting GPS points using QGIS using paths

- Different analytical tools produced different track records.
- Some paths and access points are more used than others.
- OSM GPS tracks can be used to monitor activities in greenspace.
- **Issues:** (1) No control over how many tracks available, (2) cleaning and processing the data are challenging.
Geospatial Data (summary)

- A lot of open, free, easily accessible data sources.
- Platform such as Google Earth Engine, OpenStreetMap have wide verity of Big Geo-data. GEE for LST: https://code.earthengine.google.com/229c64e5d3ea6c34af203ea2b1aeaeb4?noload=true
- Analytical tools such as QGIS, ArcGIS, R-packages, GDAL, GRASS providing opportunities to analyse Geospatial data with ease.
- Too much data! Need to be careful about using the appropriate data (e.g. resolution), scale and tools based on purpose! (will discuss more in Part-4)

The Earth Engine Public Data Catalog

- > 200 public datasets
- > 5 million images
- > 4000 new images every day
- > 5 petabytes of data

Sources:
- https://geohackweek.github.io/GoogleEarthEngine/01-introduction

Source: Muenchow et al., (2017)
Part -3: Examples of Geospatial model coupling
What all these **Geospatial data, and tools can do** in terms of decision making?

- Geospatial modelling provides the opportunity to **integrate multiple models** (e.g., earth system model, pollution) together.
- A **multidisciplinary modelling approach**.

---

**Spatial Data**:
- Capture
- Process
- Manage
- Relational database

**Model coupling for multidisciplinary modelling within GIS framework**

**Modelling**
- Model -1 (e.g. Air pollution, greenspace)
- Model -2 (e.g. Exposure)
- Model -3 (e.g. sustainability, health, prediction)

**Outputs**
- Mapping
- Assessing impact
- Reporting
- Decision making
Transport is a major determinant of global carbon emission, and it is also a major source of air pollution and related health impact (Woodcock et al., 2009).

Traffic related carbon emissions correlate with local available bio-capacity of carbon sequestration.

Can we combine two components (1) traffic carbon emission, and (2) local bio-productivity to come up a sustainability rating tool?
Example study 1: Modelling transportation sustainability (Cont...)

\[ E_i = \sum_{j=1}^{n} \sum_{k=1}^{n} EF_{ijk} A_{jk} \]

Where,

- \( i \) = Type of a pollutant (in this case \( \text{CO}_2 \))
- \( j \) = Fuels consumed (e.g. CNG, Gasoline)
- \( k \) = Emitting Vehicular type (Volume survey)
- \( E_i \) = Emissions from pollutant
- \( EF_{ijk} \) = Emission Factor (g/km)
- \( A_{jk} \) = Activity level for each pollutant source.

\[ BC = \sum_{i=1}^{n} A_{ri} * YF_i * EQF_i \]

where,

- \( BC \) = Bio-capacity (in global hectare, gha)
- \( A_{ri} \) = Area of \( i \) land use type (hectare)
- \( YF_i \) = Yield factor \( i \) type land use type (ratio of national yield world average yield)
- \( EQF_i \) = Equivalency factor for \( i \) type land use type
Example study 1: Modelling transportation sustainability (Cont...)

- Ten studied nodes
- Critical locations on the transport network.
Geospatial Model coupling

Example study 1: Modelling transportation sustainability (Cont...)

Spatially explicit estimated CO$_2$ emission

Remote sensing based land use classification
**Geospatial Model coupling**

**Example study 1: Modelling transportation sustainability (Cont...)**

### Table 5

<table>
<thead>
<tr>
<th>Area</th>
<th>Carbon Uptake Land (gha)</th>
<th>Bio-capacity Area (gha)</th>
<th>EBI</th>
<th>EBS</th>
<th>Color Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area I (Mirpur 10)</td>
<td>785.20</td>
<td>269.43</td>
<td>0.343</td>
<td>3</td>
<td>Orange</td>
</tr>
<tr>
<td>Area II (Mog bazaar)</td>
<td>987.08</td>
<td>298.06</td>
<td>0.302</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area III (Motijheel)</td>
<td>1269.36</td>
<td>278.20</td>
<td>0.219</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area IV (Gulshan 1)</td>
<td>1398.43</td>
<td>242.60</td>
<td>0.173</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area V (Shymoli)</td>
<td>1432.89</td>
<td>233.91</td>
<td>0.163</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area VI (Technical Morh)</td>
<td>1477.91</td>
<td>217.92</td>
<td>0.147</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area VII (Jatrabari)</td>
<td>1779.99</td>
<td>335.08</td>
<td>0.188</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area VIII (Mohakhali)</td>
<td>1868.61</td>
<td>317.41</td>
<td>0.170</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area IX (Science lab)</td>
<td>2363.18</td>
<td>285.57</td>
<td>0.121</td>
<td>4</td>
<td>Red</td>
</tr>
<tr>
<td>Area X (Farm gate)</td>
<td>2440.20</td>
<td>289.00</td>
<td>0.118</td>
<td>4</td>
<td>Red</td>
</tr>
</tbody>
</table>

- Emission Bio-capacity Index (EBI) = Carbon Uptake land / Bio-capacity
- Values over One (1) indicate full sequestration of CO₂ with the local bio-capacity. Expressed in four color rating; **Red, Orange, Yellow, Green**.
- 9 nodes indicated rating: “**Red**”, implying the CO₂ emission is beyond the capacity to local bio-productive areas to offset the impact.
- **Main reasons**: Increased motorized traffic volume, poor signal system, low facilitation for non-motorized vehicles, and overall low availability of greenspace.
Investigation of the likelihood of green infrastructure (GI) enhancement along linear waterways or on derelict sites (DS) using machine learning

S.M. Labib

School of Environment, Education and Development (SEED), University of Manchester, Arthur Lewis Building (1st Floor); Oxford Road; Manchester; M13 9PL, UK

- **Green Infrastructure** (e.g., greenspace, blue space) is associated with ecosystem services and health in urban areas (Tzoulas et al., 2007).

- Increased presser on urban land use resulted in loss of GI in cities.

- Can we model what would be future scenarios of GI (along waterways or existing derelict sites) based on previous trends, applying machine learning models?

- Can we compare ML models with traditional regression based models (i.e., logistic regression)?
Example study 2: Modelling Green infrastructure using ML (Cont...)

- **System 1**: Modelled or spatial data
- **System 2**: Modelling
- **System 3**: Prediction using trained model
Geospatial Model coupling

Example study 2: Modelling Green infrastructure using ML (Cont...)

Training sites (3916 along waterways, 866 derelict sites)

Prediction sites (150 along waterways, 112 derelict sites)
Geospatial Model coupling

Example study 2: Modelling Green infrastructure using ML (Cont…)

Input data from different spatial data sources, and modelled NO2 data
Prediction for Waterway corridor plots

ANFIS Prediction; RMSE: 0.29; 79.3% green
ANN Prediction; RMSE: 0.28; 80.7% green
Logistic Regression Prediction; RMSE: 0.36; 74% green
Prediction for Derelict plots

- **ANFIS Prediction; RMSE: 0.285; 61.6% green**
- **ANN Prediction; RMSE: 0.23; 53.6% green**
- **Logistic Regression Prediction; RMSE: 0.35; 34.8% green**

- **Derelict sites** are more likely to become **grey areas/buildings**, where water ways corridors plots are more likely to remain or become green areas.
- **ML models** unable to explain the importance or significance of the input variables
- Logistic regression models indicated, **site size, population density and air pollution** are significantly associated with green transformation likelihood.
• Modelling **approaches are transferable**; can be applied in different studies, such as built environment-health, air pollution-health studies

• Different spatial and non-spatial **data can be integrated** within the modelling environment.

• Emerging algorithms are being introduced/integrated frequently.
Part -4: Spatial dimensions of greenspace and health research-current practice
Spatial dimensions of the influence of urban green-blue spaces on human health: A systematic review

S.M. Labib*, Sarah Lindley, Jonny J. Huck

Department of Geography, School of Environment, Education and Development (SEED), University of Manchester, Arthur Lewis Building (1st Floor), Oxford Road, Manchester, M13 9PL, UK

- identify the different data, scales and geospatial methods utilised in studying greenspace and its relation to human health in urban areas;

- investigate how results vary (e.g., significant vs insignificant, positive vs negative) according to the type of association between greenspace and health indicators and their relation to spatial data and methods; and

- identify the limitations and prospects of spatial data and analytics in representing and associating greenspace and human health.

doi: https://doi.org/10.1016/j.envres.2019.108869
Spatial Dimensions greenhouse & health

General characteristics of the studies

(a) Number of Studies by Year

(c) Number of Studies by Health Focus

(b) Distribution of Studies by Region

Legend:
- 1 - 3
- 4 - 5
- 6 - 10
- 11 - 30

Number of Studies: 59, 18, 16

Regions: North America, Europe, Asia, Australia, South America, Africa

Total Studies: 36, 30, 16
Spatial Dimensions: greenspace & health

Spatial Scale

- Commonly used scales: **body, neighbourhood and City/districts**

- Neighborhood: (1) *egocentric* (e.g., a buffer around the home location) or (2) *allocentric* (e.g., using a pre-defined administrative unit)

- Majority of the studies focused on **ego-centric neighborhood**, applying different buffer distances (e.g., 400, 500, 800 m)
Commonly used greenspace metrics:

- **Greenspace metrics:** Land use land cover (n = 47), NDVI/EVI/SAVI (n = 36), Canopy coverage (n = 5), Street view images (n = 3), 3D viewshed (n = 3).

- Land use and Land cover data often collected at **large spatial scale** (e.g., 1:100,000); CORINE, Urban Atlas data (minimum greenspace size 25ha).

- NDVI or satellite image indices often are estimated from **Low spatial resolution satellite**, mostly Landsat (30m), and MODIS (250m).

- **Street view data are emerging**, only available along streets.
Spatial Dimensions: greenspace & health

Spatially explicit greenspace exposure assessment

- **Availability** of greenspace or greenness in different neighbourhoods (e.g., percentage, numbers, mean NDVI, and area/size). Most common (n = 75).

- **Accessibility** to greenspace from home (e.g., numbers of accessible parcels, distance to parcels) (n = 48). Measured using both shortest distance, and fixed distance (e.g., 400m).

- **Visibility** of greenspace while travelling or around the home. **Least studied** (n = 6).

- Most studies use **proximity, and overly functions in ArcGIS/QGIS**.
Spatial Dimensions  greenspace & health

Analytical approaches and key results

- A mix of subjective (e.g., self reported, GHQ12, SF36) and objective (e.g., anthropometric information, GPS tracking) health indicators (e.g., BMI, MVPA).

- Most studies based on statistical modelling (e.g., logistic, linear regressions) and correlation analysis. Very few applied spatial models (e.g., regression with lag).

- Majority of the studies found positive associations at each scale. Mixed or insignificant associations also observed at all scales.

- Neighbourhood scale has more variations in study results, as it is most commonly used, and there are a lot of variations in conceptualising neighbourhood (e.g., different buffer distances).
Majority of the studies found positive associations between health greenspace exposure.

Mixed associations and insignificant associations observed depending on how the exposure measured. Such as availability within 400m vs 1600m; the resolution of spatial data (MODIS vs. Landsat); shortest distance vs. fixed distance.

All visibility exposure studies found significant positive associations.

Absence of integrated approach of modelling exposure. Depends on different pathways.
Spatial Dimensions greenspace & health

#Issue-1 Scale of analysis, distances, and MAUP

- **Spatial unit of aggregation and analysis** is a major concern. It influences both measurements and associations.
- Different buffering approaches (e.g. Euclidian, Network), and administrative units produced **different exposure areas**, and spatial aggregation of model inputs.
- **Physical health focus studies** usually use larger distance than mental health.
Spatial Dimensions greenspace & health

#Issue-1 Scale of analysis, distances, and MAUP (cont…)

- Varying distances, spatial units, and buffering approaches result in **Modifiable Areal Unit problem - MAUP** (scale effect/ aggregation, zone effect).

- Aggregating into **larger spatial scale reduce variance**, cause inconsistency in the model.

- Studies used **larger buffers** to measure greenspace exposure usually found significance associations, but **effect sizes become inconsistent**, as covariance among variables affected.

- Zoning of the exposure areas also effect the variance, and hence influence the associations.

Source: Dark and Bram, (2007)

**Effects of Aggregation**

<table>
<thead>
<tr>
<th></th>
<th>a.</th>
<th>b.</th>
<th>c.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3.5</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>3.75</td>
</tr>
</tbody>
</table>

\[ \bar{x} = 3.75 \]
\[ \delta^2 = 2.60 \]

\[ \bar{x} = 3.75 \]
\[ \delta^2 = 0.50 \]

\[ \bar{x} = 3.75 \]
\[ \delta^2 = 0.00 \]

**Effects of Zoning Systems**

<table>
<thead>
<tr>
<th></th>
<th>d.</th>
<th>e.</th>
<th>f.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5</td>
<td>2.75</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>4.75</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>4.5</td>
<td>4.5</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{x} = 3.75 \]
\[ \delta^2 = 0.93 \]

\[ \bar{x} = 3.75 \]
\[ \delta^2 = 1.04 \]

\[ \bar{x} = 3.17 \]
\[ \delta^2 = 2.11 \]
Some potential solutions:

- Select an unit of analysis, or buffer distance that do not cause over aggregation of exposure or health variables. Need sensitivity analysis. [My upcoming paper detailed with this issue]

- Use a weighted/fuzzy distance approach, when do not know what distances more appropriate (Chaix et al., 2009), for which health focus.

- Use activity space to determine more realistic exposure area. Smith et al., (2019) detailed some state-of-the art approaches in activity space delineation.
# Issue-2 Resolution of images and data capturing scale

- Resolution of the metrics of greenness can cause misclassification of greenness, and result in under or over estimation of exposure.

- Low spatial resolution could also result in insignificant/ mixed association with health outcomes (also Reid et al., 2018).

- Scale of analysis/ aggregation area sensitive to data resolution.

**Some potential solutions:**

- Use the best resolution data available, currently Sentinel-2 is the better free option (Part 2, case study 1)

- Select an aggregation unit/ exposure area/scale that does not over aggregate already misclassified exposure. [My upcoming paper investigated this for satellite images]
Spatial Dimensions  greenspace & health

#Issue-3 Spatial autocorrelation

- All spatial data usually have some autocorrelation, mostly positive.
- Autocorrelation among observations can be introduced with overlapping exposure areas.
- Auto-correlated variables usually has less information, reduced effective sample size, and vulnerable to Type-I error, when using in a non-spatial modelling approach (e.g. linear regression).
- Spatial autocorrelation observed in few greenspace and health studies, most studies did not checked.

Some potential solutions:
- Test autocorrelation (e.g., Moran 'I)
- Apply spatial smoothing, or randomization.
- Apply spatially explicit regression (e.g., Geographically weighted regression, Bayesian spatial model), and test application of ML algorithms (Part 3, example 2).

(a) Positive spatial autocorrelation. (b) Spatial randomness. (c) Negative spatial autocorrelation
(Source: Fortin and Dale, 2009)
Take home message

Part-1
• Spatial data and methods are integrated in environmental epidemiological studies
• Environmental exposure assessment frequently dependent on spatial methods.

Part-2
• A lot of spatial data available, can be used in different epidemiological studies.
• Free, open and easy access to big-spatial data via platforms like Google earth engine, OpenstreetMap. A lot of open access analytical tools available.

Part-3
• Spatial modelling framework provide opportunities to integrate multiple data, and models
• Adopting new algorithms allowing robust modelling
• Transferable modelling approach

Part-4
• Applying spatial data, methods require careful attention in selecting data types, scale of analysis, and methods.
• Scale, resolution, MAUP, and autocorrelation can influence the associations among variables.
• Fine resolution data, appropriate scale, and spatially explicit modelling should be used environmental epidemiological studies.
Thank you...
Any Questions!

sm.labib@manchester.ac.uk

@smlabib
References

Barton, H. and Grant, M., 2006. A health map for the local human habitat. The journal for the royal society for the promotion of health, 126(6), pp.252-253.


