Chapter 3
Spatial Variability and Data Analysis in Urban Soils

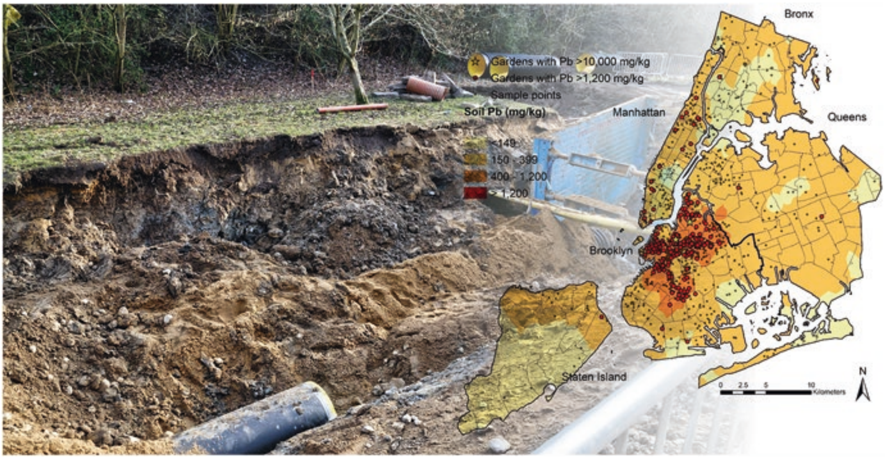
Andrew W. Rate

Abstract Urban soils are likely to be even more variable than soils in other environments, due to the inherent heterogeneity of urban environments which intensifies natural soil variability. This chapter examines soil variability related to urbanisation at several different spatial scales, from regional phenomena to differences observed on the scale of individual soil profiles. Soil sampling strategies and designs are described, with discussion of the issues related to sampling density, sample numbers, the geometric arrangement of sampling locations, and 'hotspot' detection. We present methods for qualitative and quantitative analysis of soil spatial data using maps, spatial autocorrelation analysis, and variograms and kriging. Basic but rigorous statistical methods are described in the context of soils and compositional data, including comparisons, relationships, and multivariate techniques.

Keywords Variability · Urban soils · Spatial scales · Spatial analysis · Spatial autocorrelation · Kriging · Statistics · Data analysis

A. W. Rate (✉)
School of Agriculture and Environment, University of Western Australia,
Crawley, WA, Australia
e-mail: andrew.rate@uwa.edu.au

A. W. Rate (ed.), *Urban Soils*, Progress in Soil Science,
https://doi.org/10.1007/978-3-030-87316-5_3



What you could learn from this chapter:

- That soils are naturally variable, and whether urban centres create additional sources of soil variability. You'll also consider the spatial scales over which soils vary and how these are related to urban environments.
- How to sample soils in urban environments – there are several strategies, and we will discuss the reasons why we choose certain depths, combinations of sampling locations, how many samples we should take, and how far apart.
- How we conduct spatial data analysis to investigate the relationships between soil samples from different locations and the specific spatial statistical techniques we use.
- How we apply common statistical methods to describe, explore, and assess urban soil data.

3.1 Soil Variability in Urban Environments

Urban environments are far from uniform; cities are characterised by extreme variability, reflecting the range of types, intensity, and timescales of human modification of the natural environment (Grimm et al. 2000; Pickett et al. 2001). Not surprisingly, this general heterogeneity of urban environments is also present in urban soil environments. Prior to human interference, natural soils already showed significant spatial variation due to differences in soil parent material and other soil-forming factors (see Chap. 2). This pattern of natural soil variability has, superimposed upon it, the imprint of diverse human activity, creating an even more variable soil landscape.

3.1.1 Cities and Regional Soil Variability

There is some evidence from continental-scale geochemical surveys (Cicchella et al. 2015; Mann et al. 2015) that urbanised areas can be distinguished from non-urbanised land based on the concentration of some (potential contaminant) elements such as lead (Pb), with an example shown in Fig. 3.1. This distinction between urban and non-urban areas is detectable despite the variable background from differences in the chemical composition of parent materials. The geochemical signature of urbanisation extends to peri-urban areas (Cicchella et al. 2015). The anthroposequences discussed in Chap. 1 and elaborated on in the work of Richard Pouyat and others (e.g. Pouyat and McDonnell 1991; Pouyat et al. 2007; Pouyat et al. 2008) also provide evidence that soils in cities may be distinguished from their rural counterparts on a regional scale.

3.1.2 Soil Variability at the Scale of Cities

The next largest scale of soil variability in urban environments is constrained by the dimensions of an individual city, taking into account the effects of the urban centre on the surrounding peri-urban and non-urbanised land.

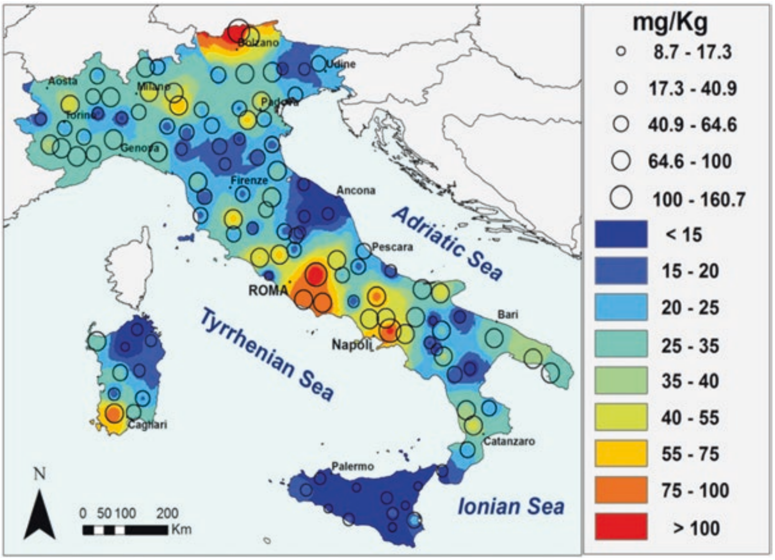


Fig. 3.1 Lead (Pb) concentrations in Italian agricultural surface soils, showing the effect of the urban centres of Roma and Napoli (from Cicchella et al. (2015) and used with permission from Elsevier)

An important parameter for the spatial distribution of soils in cities is the amount of soil remaining exposed after urban development. This can be assessed using conventional ground-based surveying techniques or by remote sensing and image analysis (Wu and Murray 2003; Wu 2004). The remote sensing techniques commonly categorise land into vegetation, impervious, or soil categories (the V–I–S model). Urban catchments can range from 5 to 100% impervious cover, depending on the density of infrastructure (Fletcher et al. 2004). Assuming that the vegetation component of the remaining pervious cover represents plants growing in soil, then urban soil cover may range from approximately 95% to 0%. Whole-city soil cover ranges from 10 to 35% for smaller cities (pop. < 100,000; Bauer et al. 2008) and up to at least 60% in larger cities (Zhu et al. 2017). The proportion of soil cover in urban environments is itself highly variable and shows gradients related to distance from urban centres and age of urban development (Powell et al. 2007). Soil cover tends to be lowest in urban centres and to decrease as the time since development increases.

Other urban soil properties also show systematic variation on whole-city scales. Johnson and Ander (2008) reviewed multiple studies of the spatial distribution of trace elements in urban environments, explaining that such studies have multiple objectives, including collection of baseline data, and identifying contaminated areas and their sources and risks. In many cities, the concentrations of some contaminants are greatest in the metropolitan centre (often the oldest area in the city), and lower concentrations occur at greater distance from the urban centre (e.g. Pb concentrations in surface soils in Pueblo, Colorado (Diawara et al. 2006); see Fig. 3.2). This implies a cumulative and ongoing input of contaminants, relating to more diffuse sources such as road traffic, construction, etc. In other urban environments, the concentrations are greatest and decrease with distance, from a recognisable source of contamination. For example, lead (Pb) concentrations in surface soil in the city of Mount Isa, Queensland, Australia, where Pb and Cu are mined, decrease from the mine and smelters in the west towards the urban area in the east, despite the dominant SE wind direction (Taylor et al. 2010).

3.1.3 Soil Variability at the Locality or Site Scale

The spatial distribution of potential contaminants (usually trace metals) in urban soils may simply represent variations in the concentrations of the same elements in the soil parent materials (e.g. Co and Mn concentrations measured by Gong et al. 2010) (see Fig. 3.3). In such cases, the potential contaminants are termed *geogenic*, emphasising that their origin is not from human activity. Depending on the size of the urban area, geogenic variations in soil properties vary on a similar scale to whole-city anthropogenic variations. For other soil properties (e.g. contaminant concentrations), the variability is attributable to human modifications of the soil environment. These modifications include excavation, dumping, or construction and point or diffuse sources of contamination such as vehicle traffic or industrial emissions. For example, the high concentrations of Pb and Zn in urban centre soils

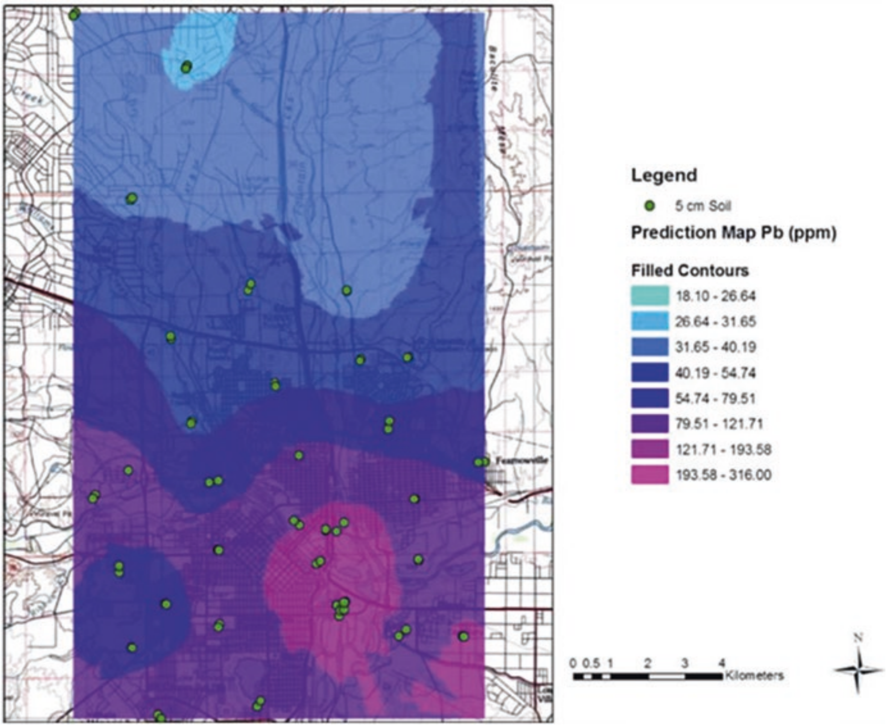


Fig. 3.2 Spatial distribution of lead (Pb) in surface soil (0–5 cm) sampled in Pueblo, Colorado, USA (from Diawara et al. (2006); used with permission from Springer)

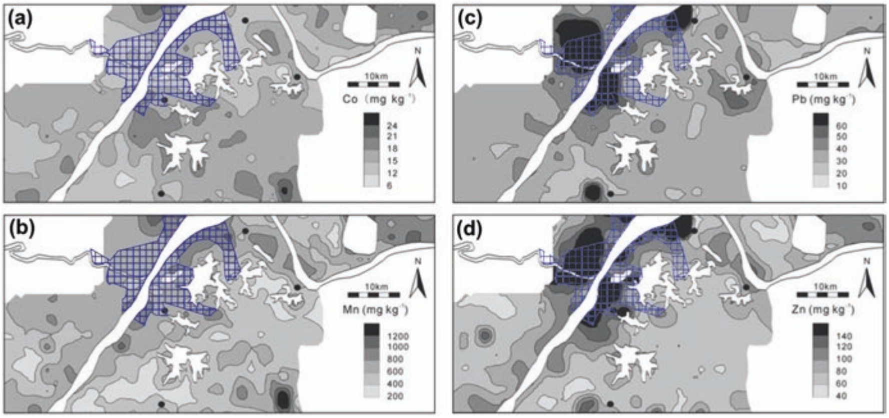


Fig. 3.3 Concentration distributions of (**a**) cobalt, Co; (**b**) manganese, Mn; (**c**) lead, Pb; and (**d**) zinc, Zn, from a gridded sampling design in the area in and around Wuhan city, Hubei, China (blue cross-hatching ▦ shows urban area). Cobalt and manganese (**a** and **b**) show a geogenic distribution, whereas lead and zinc (c and d) show accumulation in the urban area related to anthropogenic additions. Redrawn from Gong et al. (2010); used with permission from Springer

in Wuhan, China, were attributed by Gong et al. (2010) to domestic heating and road traffic.

In practice, urban soil investigations are more often conducted at the scale of specific sites, for example, to obtain site-specific data for environmental impact assessment (e.g. Carr et al. 2008). Figure 3.4 shows a typical pattern of spatial variability in soil properties for urban parkland soils, where the soil properties reflect a combination of both geogenic variation (some of the parkland occupies former lake beds) and anthropogenic modification (horticulture, landfill, construction, and demolition wastes). For example, zones of low pH in Fig. 3.4 are probably due to drainage of lake sediments and subsequent formation of acid sulphate soils; high pH probably reflects the use and/or disposal of limestone-based products such as construction cement. High EC (a measure of soluble salts in soils) could be caused by release of dissolved ions by acid sulphate oxidation or demolition of a building containing gypsum-based materials (e.g. wall and ceiling panels) in the south of the study site.

Soil variability at the site scale is also shown more simply in concentration-distance relationships, for example, along transects (which are analogous to the anthroposequences described for city- or regional-scale transects by Pouyat et al. (2008)). Figure 3.5 presents two examples of such variability, showing gradients away from likely contaminant sources (a major roadway and a coal-fired power station site). In both cases, it is likely that contaminants are carried from their sources by wind (as aerosols) and deposited on soil surfaces in decreasing amounts as distance from the source increases. Transport of material from roads to soils by surface flow of water is also possible.

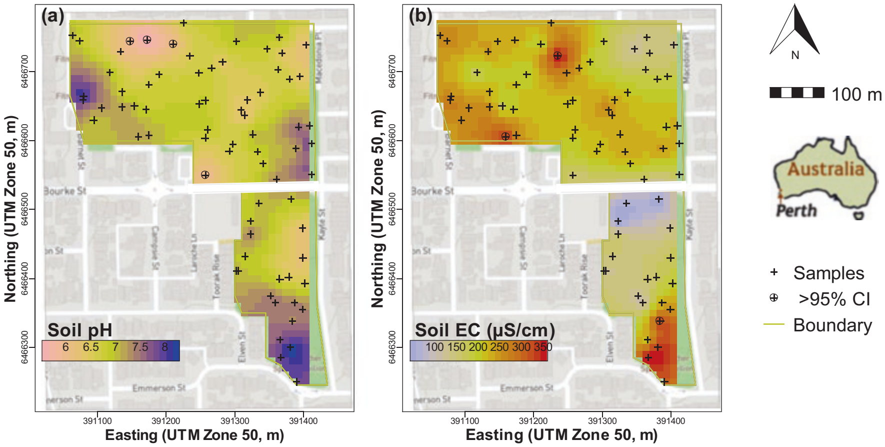


Fig. 3.4 Spatial variability of (**a**) surface soil pH and (**b**) electrical conductivity (EC) of a 1:5 soil/water suspension at Charles Veryard and Smith's Lake Reserves, urban parklands in Perth, Western Australia. Maps generated in R with the packages 'OpenStreetMap' (Fellows 2019) and 'geoR' (Ribeiro Jr. et al. 2020), and using kriging with exponential variogram models (graphic by Andrew W. Rate)

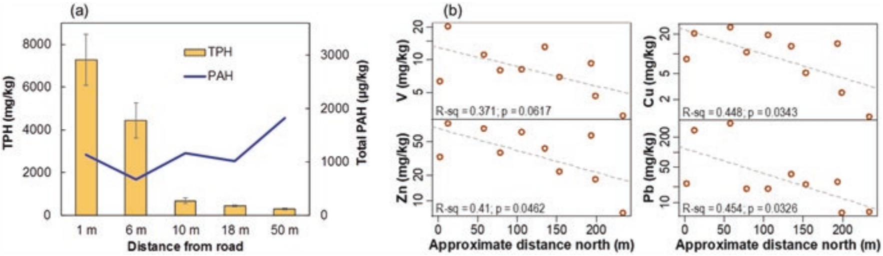


Fig. 3.5 Concentration-distance gradients for (**a**) total petroleum hydrocarbons (TPH) and polycyclic aromatic hydrocarbons (PAH) in urban roadside soils (redrawn from Nikolaeva et al. 2017, and used with permission from Springer) and (**b**) metals V, Zn, Cu, and Pb in urban parkland soils north of a decommissioned coal-fired power station (Banks Reserve, Perth, Western Australia; graphic by Andrew W. Rate)

3.1.4 Contamination Hotspots

Some urban events, such as chemical spills, can produce soil variability on a very small scale. Relatively small (metre-scale) areas of high concentration are commonly termed ‘hotspots’. An example of a contamination hotspot at this scale is shown for an urban public open space in Fig. 3.6. In the example shown in Fig. 3.6, the contamination with several metals in the hotspot most likely represents disposal of contaminated sediments from the adjacent open stormwater drain onto the soil surface during drain maintenance. The hotspot was identified by sampling soils on transects along expected contamination gradients, rather than grid sampling, so in a sense the hotspot was identified ‘accidentally’!

The term ‘hotspot’ is not reserved for metre-scale areas, however; Nriagu (1988) argues that urban areas themselves represent hotspots on regional, continental, or even global scales. Similarly, Li et al. (2004) identify district-sized hotspots within the city of Hong Kong. A better definition of a hotspot, then, is an area of high concentration of contaminants in soil that is localised, or small, relative to the scale of observation.

3.1.5 Soil Variability with Depth

Soil properties can vary substantially with depth, that is, in the vertical rather than horizontal dimension which has been the focus of the sections above. For some soil properties, such as organic carbon content, the source of the variation is natural pedogenesis. The process of soil formation results in greater organic carbon content in a soil’s surface horizon(s), since most additions of organic materials (litter fall, excretion) and losses (microbial decomposition) occur in the surface soil (Coleman et al. 2017). Natural pedogenesis may also result in enrichment of some soil components at greater depth, such as clay or iron content, due to the complex

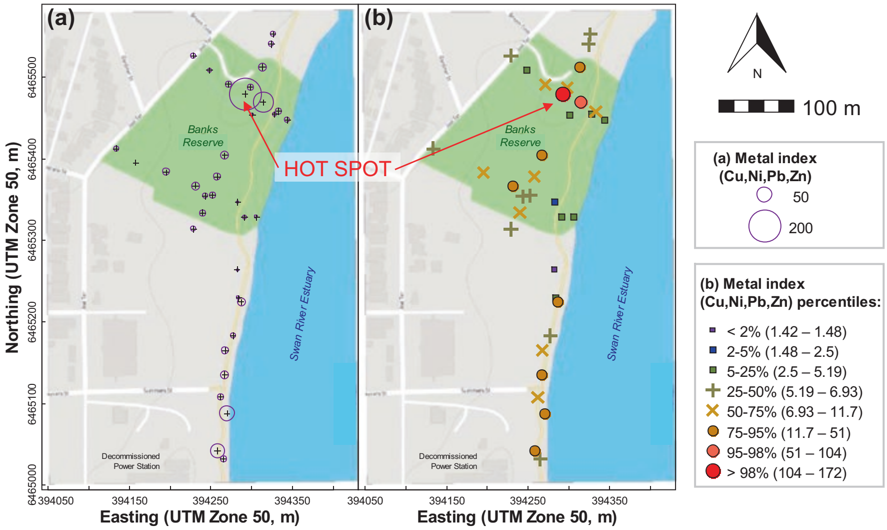


Fig. 3.6 A soil contamination hotspot (labelled) shown by the spatial distribution of a metal contamination index (an additive index of normalised Cu, Ni, Pb, and Zn concentrations – contamination indices will be explained in Chap. 6). Data are shown with two representation styles for point spatial data: (**a**) area-proportional 'bubbles' and (**b**) symbols showing percentile categories

interactions of soil and hydrological processes. General soil science texts, such as those by Schaetzl and Anderson (2005) and White (2006), discuss natural pedogenic processes resulting in depth variation in much more detail.

In urban soils, soil properties do not always vary with depth in straightforward ways. A huge variety of human modifications, such as additions of new material (and often burial of existing soils), excavation and refilling, and contamination, are commonly superimposed upon natural pedogenic processes. Figure 3.7 shows some depth profiles for soil properties in a highly disturbed urban area.

In some cases the concentration vs. depth relationships (referred to as depth profiles) in urban soils, even for contaminants, are closely related to depth profiles for pedogenic features. For example, metal contaminant concentrations in soil adjoining highways in Cincinnati, USA, decreased with increasing depth (Turer et al. 2001) but were shown to be more closely related to soil organic matter content than to other predictors. A similar control by soil organic matter was deduced for phthalate and BPA contaminants (organic compounds used as plasticisers) in urban and other soil environments (Tran et al. 2015) and for polycyclic aromatic hydrocarbons (PAH) in soils in Beijing, China (Bu et al. 2009). The relationship of contaminants to a pedogenic parameter such as soil organic matter content demonstrates the ability of organic matter to act as a sink for added contaminants. Soil clays or iron oxides may also provide contaminant sinks (e.g. Rate 2018).

For biological parameters, the rule of thumb that biological activity decreases as soil depth increases may not apply in all urban soil environments. For example, Zhao et al. (2012) found that soil microbial biomass (measured as carbon or

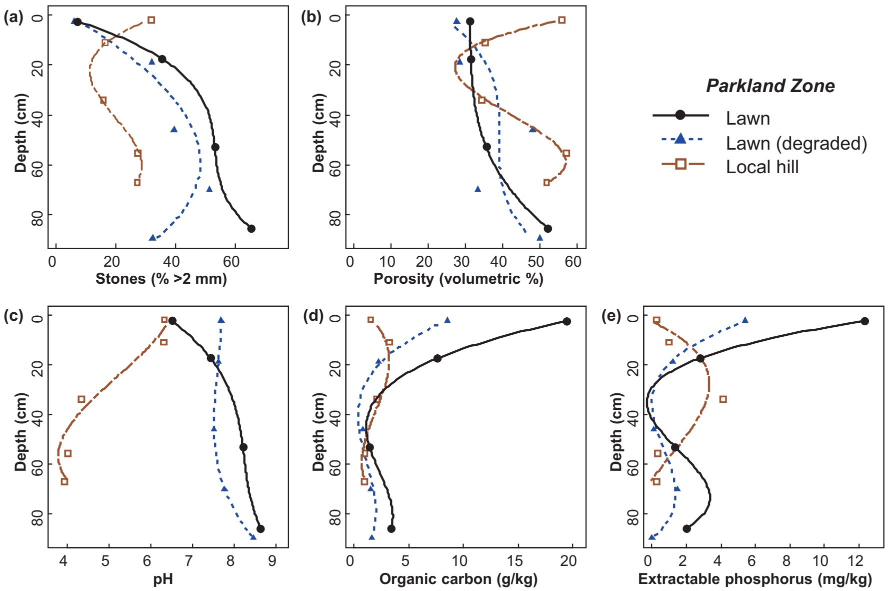


Fig. 3.7 Variation of physical and chemical soil properties with depth, in soils from an urban parkland in Hong Kong (redrawn by Andrew W. Rate from data tabulated in Jim 1998). Smooth lines are cubic regressions intended only as visual guides. The high stone contents (**a**) reflect the use of fill material, with many low porosity values (**b**) reflecting compaction by foot traffic. High pH (**c**), except in the relatively natural local hill soil, is related to use of cement-based materials. The generally low organic carbon (**d**) and extractable P (**e**) contents reflect the young age of the soil materials, with concentration peaks at depth possibly indicating burial of pre-existing soil

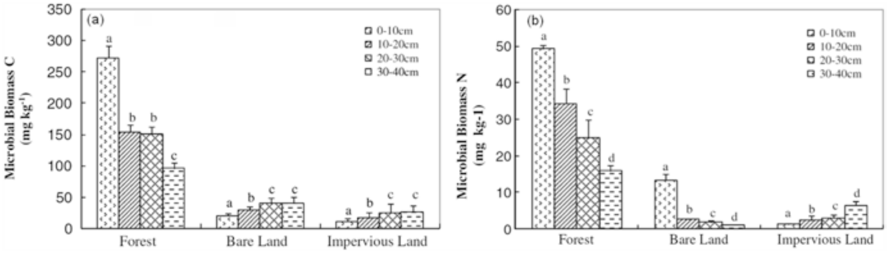


Fig. 3.8 Variation with depth of soil microbial biomass carbon (**a**) and nitrogen (**b**) in different land cover types in urban Beijing, China (from Zhao et al. 2012 and used with permission from Springer)

nitrogen) could decrease or increase with increasing depth, in urban soils of Beijing, China, depending on the measurement or land cover type (Fig. 3.8). In this study, microbial biomass C or N was greater in deeper soil samples under impervious cover (Zhao et al. 2012).

3.2 Measurement and Description of Soil Variability

The previous sections in this chapter described the scales of soil variability in urban systems and provided some explanations for this variability. Soil variability is, logically, more pronounced in urban soil environments, due to the combination of a wide range of natural and human processes which affect soil properties and formation. The following sections will therefore deal with how soils can be characterised in highly variable urban environments and how the variability can be both accounted for, and described, by suitable sampling and statistical methods. In the next chapter, we will also address some of the implications of soil heterogeneity for land utilisation in cities.

The way in which urban soils are sampled is guided by the purpose(s) of the sampling program. Johnson and Ander (2008) review multiple urban geochemical surveys and identify several objectives for mapping soil properties (especially contaminant concentrations) in urban soil environments. These objectives include determining the current (baseline) status of an urban environment; regulatory drivers, such as fulfilling the requirements for environmental impact assessment for development; identifying and locating polluted urban areas; identifying (potential) contaminant sources, including separating geogenic and anthropogenic sources; and assessing risks to other urban environmental compartments such as water bodies and human health (Johnson and Ander 2008).

3.2.1 Sampling Depth

In all of the strategies outlined in Sect. 3.2.2 for generating a two-dimensional spatial arrangement of sampling points, a decision must also be made about the depth range(s) of soil to be sampled. In some protocols (e.g. Smith 2004), this is guided by the depth of a particular soil horizon, commonly the A horizon but in other studies different horizons are targeted (e.g. Reimann et al. 2008, who describe sampling of O, B, and C horizons). In soils in urban environments, horizons are commonly poorly developed or obscured by truncation or addition of material. As a result, sampling of urban soil profiles is usually done on the basis of depth increments; 0–10 cm is common for surface soils, and narrower increments are more subject to sampling variability (Johnson and Ander 2008). Depth ranges for ‘surface soils’ in the published literature generally range from 0–2 cm to 0–20 cm. When assessing an environment for contamination, many protocols (e.g. Department of Environmental Protection 2001) recommend sampling at multiple depths, and examples of studies where depthwise sampling has been performed are Turer et al. (2001), Zhang (2004), Zhao et al. (2012), and Corrò and Mozzi (2017). Some studies assume that deeper samples remain relatively uncontaminated, especially for the purposes of calculating contamination indices (e.g. Gong et al. 2008; see Chap. 6).

3.2.2 Sampling Strategies and Designs for Urban Soils

3.2.2.1 Sampling Density

The number of samples required depends on the spatial scale and objectives of the study. In their review, Johnson and Ander (2008) identify 'systematic' and 'targeted' surveys. Systematic surveys are based on hundreds to thousands of samples over entire urban or regional areas at densities of 1–4 samples per km^2. In contrast, targeted surveys have tens of samples (or fewer), within a specific urban land use, at densities of 4–50 samples per km^2. The specific case of hotspot identification discussed below has more stringent constraints on the number of samples required for identifying very localised contamination. If there is a requirement for samples to be independent, then spatial statistics in the form of variogram analysis may be required. Essentially a variogram is an analysis of the variance (in some variable, such as the concentration of a contaminant) between samples as a function of distance between the samples. In theory, sufficient distance between samples maximises the between-sample variance, meaning that samples are independent (Oliver and Webster 2014). We will look into variogram analysis, and how it is used in practice, later in this chapter. It should be said, though, that the variogram analysis requires field data – but in this context, the purpose of using the variogram is to guide the sampling required to obtain that field data! The use of variograms to inform sampling therefore requires preliminary data to be acquired, which is unlikely in practice except perhaps for studies for scientific research purposes.

3.2.2.2 Sampling for Hotspots

The general aim of identifying and locating polluted urban areas can be specified as the search for contamination hotspots. There are methods for calculating the required number of samples for hotspot identification based on rigorous statistical principles. The Western Australian Department of Environmental Protection (2001) describes such a method, shown in Box 3.1. Bugdalski et al. (2014) emphasise that if the objective of sampling is to detect hotspots, then inadequate sampling can lead to type II errors or false negatives (i.e. not all hotspots on a site are detected).

Box 3.1 Calculating sampling grid size and sample numbers for hotspot detection

The grid size, G, should be calculated using Eq. 3.1:

$$\mathbf{G = R / 0.59} \tag{3.1}$$

where: **G** = grid size of the sampling plan, in metres **R** = radius of the smallest hotspot that the sampling intends to detect, in metres **0.59** = factor derived from 95% detection probability, assuming circular hotspots

The number of sampling points **n** should then be calculated from Eq. 3.2:

$$\mathbf{n = A / G^2} \tag{3.2}$$

where.: **A** = area to be sampled, in square metres
G = grid size of the sampling pattern, from Eq. 3.1, in metres.

Of course the size of unknown hotspots (**R** in Eq. 3.1) cannot be estimated in advance with complete confidence, so there is again the possibility of type II errors, unless the estimate of **R** is biased towards lower values. As with calculation of the maximum distance between independent samples, the size of contamination hotspots will, in practice, most often be measured after sampling is completed.

Grid Sampling Sampling urban and peri-urban soils on regular grids is most commonly used over larger (city-wide or regional) scales to assess the current state, or baseline, of a soil environment. This would equate to systematic sampling in Johnson and Ander's (2008) study; the other, more common, usage of the term 'systematic sampling' is a sampling design that is based on regular intervals across a landscape, such as various types of grid (Fig. 3.9). The geometry of a grid may be rectangular. For example, Lv et al. (2015) sampled soils across an entire province in China on a rectangular 2 × 2 km grid (0.25 samples/km^2; see Fig. 3.10). A grid sampling design provides a dataset which is well-suited to statistical spatial analysis (see Oliver and Webster 2014). For some targeted soil sampling, the lesser number of samples may be more suited to a transect (essentially a one-dimensional grid) across expected gradients in soil properties.

Stratified Sampling In some urban soil environments, there is sufficient pre-existing information to classify land into categories. This information includes data such as soil types, underlying geology, geomorphological zones, previous or current land use, vegetation communities, and so on. In such cases the total sampling area can be subdivided into sub-areas, or sampling strata, with samples collected within each of the strata (USEPA 2002). An example of a quite complex stratified sampling design for the city of Kumasi, Ghana, conducted by Nero and Anning (2018) is shown in Fig. 3.10. Stratified sampling may be used to ensure that sufficient sample numbers are taken in each sub-area and therefore can be used to test hypotheses, for example, about differences in soil properties between sampling strata.

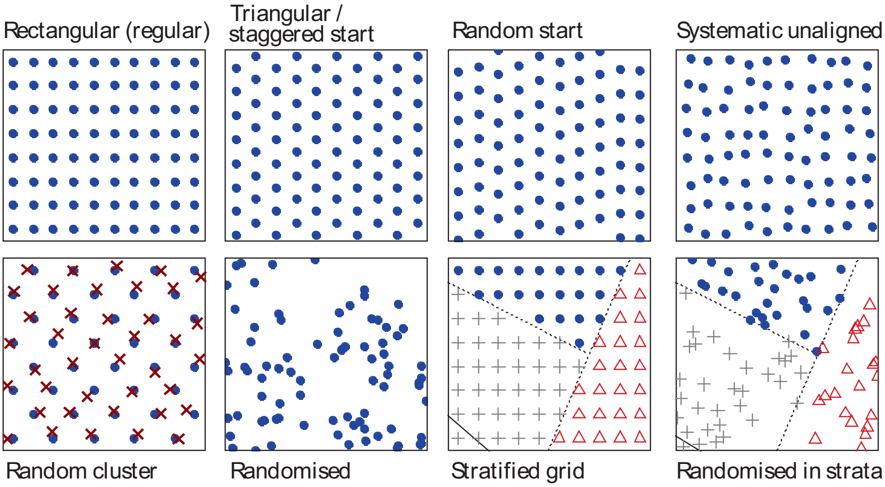


Fig. 3.9 Geometries of grid, cluster, and random sampling designs (graphic redrawn by Andrew W. Rate, based on multiple unattributed online sources)

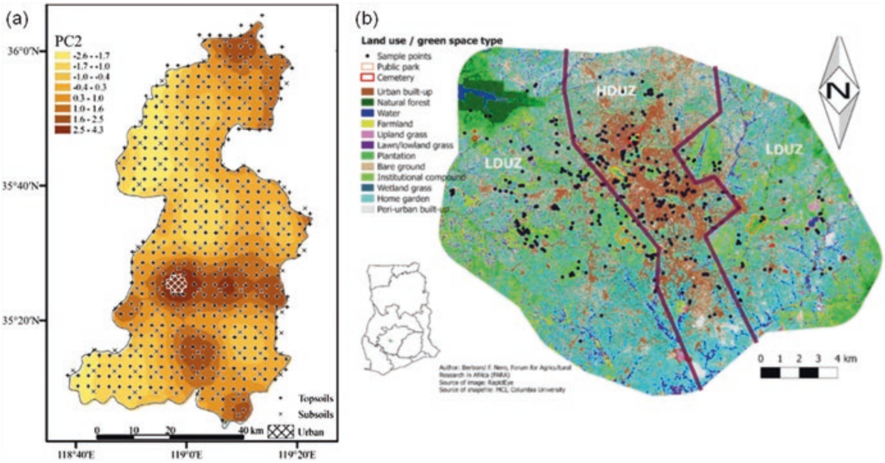


Fig. 3.10 Examples of soil sampling designs in urban and peri-urban environments
(**a**) Grid sampling Ju County, Shandong Province, China (redrawn from Lv et al. 2015; used with permission from Springer), superimposed on an interpolated map of principal component 2 (PC2, which emphasised signals from Cd, Pb, and Zn) from a multivariate statistical analysis
(**b**) Stratified sampling in the city of Kumasi, Ghana (Nero and Anning 2018; used with permission from Springer). Samples were taken from strata defined by eight unique green space types within each of two urban zones: high (HDUZ) and low density (LDUZ)

Randomised Sampling Completely randomised sampling is straightforward to implement and may be suitable for areas which have minimal variation (e.g. no hotspots USEPA 2002). Samples are theoretically statistically independent, but the

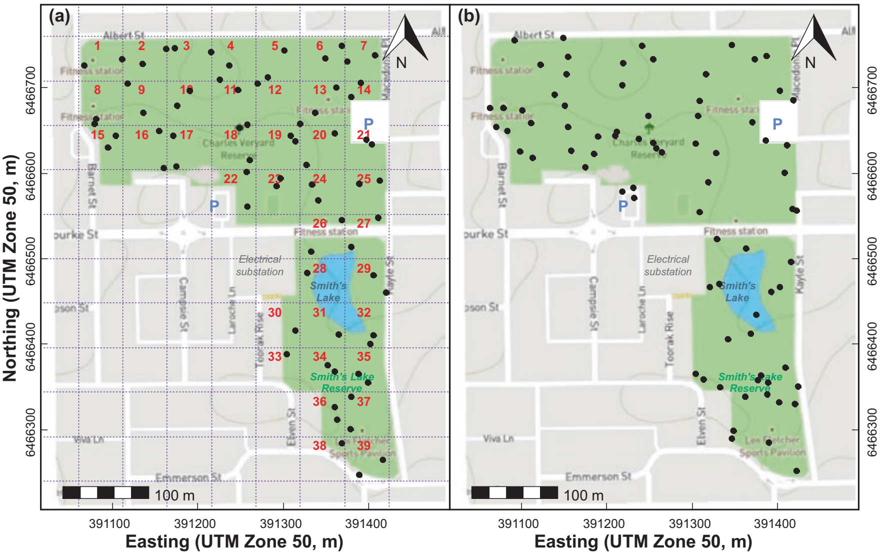


Fig. 3.11 (**a**) Random-in-grid sampling plan used to generate the data in Fig. 3.4. The grid size (52 m) was adjusted to fit the sampling area, and two samples were taken per grid square. Filled circle symbols ● show planned sampling locations with known coordinates; these were adjusted if necessary during field sampling (e.g. if a sample position was on a paved surface). (**b**) A completely randomised sampling equivalent of (a) showing large gaps in sampling area caused by uncontrolled randomisation. Other issues (e.g. samples on paved areas) can be addressed by adjustments during field sampling. Graphics by Andrew W. Rate

completely random location of sample points may mean that 'gaps' exist within the sampling area (Fig. 3.11), such that features like hotspots are missed. Since urban soils are very heterogeneous, completely randomised sampling is seldom used and is not recommended.

A variation on randomised sampling is random-in-grid sampling (Fig. 3.11), where samples are taken at randomised locations within each polygon (or stratum), or using a predetermined grid. The effect resembles cluster or systematic unaligned sampling (see Fig. 3.9) and is a compromise between the good area coverage of systematic grid sampling and the statistical independence of randomised sampling.

3.3 Analysis of Spatial Data

3.3.1 Maps

Spatial information is complex, and the most straightforward way of assessing spatial data from urban soil environments is using maps. The map itself presents the land surface as a two-dimensional representation, and other layers of information (such as land elevations, soil properties, contaminant concentrations, etc.) can be represented in different ways such as contour lines, sets of related symbols (e.g. size proportional to concentration), and so on. We have already seen Figs. 3.3 and 3.4 which show examples of contour-based information and Fig. 3.6 which has two different examples of symbol sets to show information layers on maps.

Maps are usually created in Geographic Information System (GIS) software, although more general open-source platforms such as R (R Core Team 2020) in combination with specific packages (such as those by Fellows 2019; Ribeiro Jr. et al. 2020) can provide excellent results. In addition, the QGIS package (QGIS.org 2020) is an open-source alternative to commercial GIS software and has a wide user base.

3.3.2 Spatial Autocorrelation

One of the objectives of spatial analysis is to investigate the effect of between-sample distance on soil variables. In effect, this is a test of Tobler's first law of geography (Tobler 2004), stating:

> *Everything is related to everything else,*
> *but near things are more related than distant things.*

Whether samples are spatially related or not, in terms of a particular soil property, is expressed by the spatial autocorrelation or Moran's I (Zhang et al. 2008). The Moran's I statistic is based on comparison of the values of a variable at one point with a specified number (or within a specified distance) of neighbouring points. A positive Moran's I autocorrelation suggests that locations with similar values of the variable considered tend to cluster together. A Moran's I close to zero suggests no autocorrelation (values of the variable considered are randomly located), and a larger negative Moran's I suggests that similar values of the variable considered tend to be further apart than expected from a random spatial distribution. The Moran's I statistic is tested against the null hypothesis of no spatial autocorrelation and will vary with the number of neighbouring points in the calculation, with more points giving weaker autocorrelations (Kalogirou 2019).

The basic Moran's I is a global autocorrelation, across the whole spatial dataset being analysed. The local Moran's I can also be calculated and shows the extent of significant spatial clustering of similar values (of the variable considered) around each observation. The two examples in Fig. 3.12 show local Moran's I as map symbols for the soil pH and EC data which we have already seen in Fig. 3.4. Interestingly, at the urban parkland location in Fig. 3.12, soil pH at most sampling points is spatially associated with similar values, but EC is not.

Spatial autocorrelation statistics, usually Moran's I, can be calculated using various GIS and statistical software, including several packages which add functionality to R. A useful expansion of the type of spatial autocorrelation information presented in Fig. 3.12 is local indicators of spatial association (LISA) analysis (Anselin 1995).

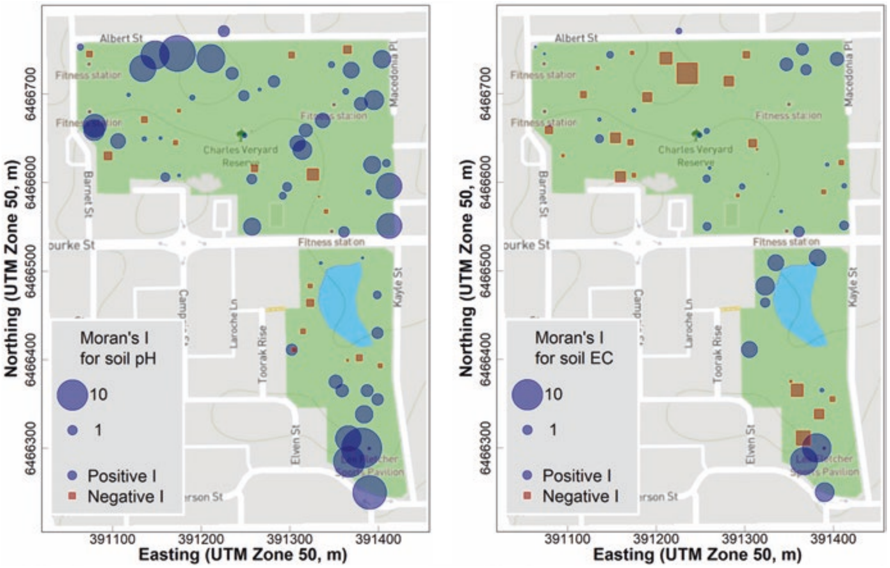


Fig. 3.12 Local Moran's I autocorrelation maps for the soil pH (left) and EC (right) data in Fig. 3.4. Larger positive values of local Moran's I imply significant spatial clustering of similar values; negative local Moran's I implies significant spatial clustering of dissimilar values. Global Moran's I (5 neighbours) for pH is positive and significant ($p \leq 0.001$,.) but for EC is near zero and $p > 0.05$ (graphics and data by Andrew W. Rate)

A LISA analysis identifies, for map points or polygons, whether significant values of local Moran's I represent association of high values of a variable with other high values, low with low, high with low, or low with high.

3.3.3 *Variograms and Kriging*

The variogram, simplistically, is the relationship between the variance between sample points and the distance separating those sample points. In many instances, it is desirable to predict a soil property at locations where samples have not been taken, and this requires some assumption(s), usually a mathematical model, about how that soil property varies with distance. This information is actually provided by the variogram. In many cases the form of the variogram relationship can be simulated adequately with a mathematical function. The variogram function can then be used to interpolate between points – a process known as kriging (after the originator of the method and pioneer of geostatistics, Professor Danie Krige). Variograms and kriging are summarised expertly by Oliver and Webster (2014) and Reimann et al. (2008). Webster and Oliver (1993) argue that at least 100 observations (and preferably more) are needed for kriging interpolation, based on variogram analysis to establish the relationship between sample points as a function of separation

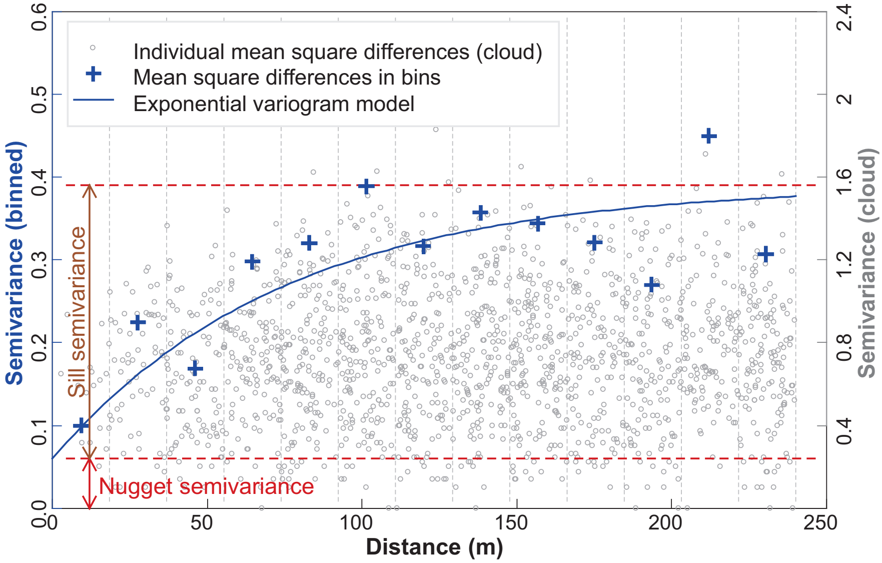


Fig. 3.13 The empirical variogram used to interpolate the soil pH in Fig. 3.4. The individual mean square differences are plotted at all possible pairwise distances between points up to the defined maximum distance (240 m, where the maximum possible pairwise distance at the site was 599 m). Vertical dashed lines represent the boundaries of the distance intervals ('bins') used to calculate the binned variogram for fitting the exponential model. The model parameters were nugget = 0.06, sill = 0.33, and range = 75 m (practical range = 225 m)

distance. The US Environmental Protection Agency (2002) reports that, for variograms and kriging, stratified sampling can have a lower sample number requirement than a simple grid but that kriging accuracy is similar for all sampling designs.

Figure 3.13 shows some of the key concepts of variogram analysis. There is some semivariance that exists even for very closely spaced samples, and this is called the 'nugget'. This semivariance increases with increasing distance between samples to a limiting value called the 'sill'. At some distance there is no increase in semivariance (which then approaches the variance of the complete dataset), and this distance is called the 'range', the value of which depends on the mathematical model used to describe the semivariance-distance relationship. The 'practical range', the distance at which samples are independent, is related to the model range by a factor dependent on the model equation.

Kriging and the associated variogram analysis can be very subjective in practice (Bohling 2005). Real soil data do not behave in an ideal fashion (see the scatter of binned points in Fig. 3.13), and there is no systematic way to make the choices of:

- The number of inter-sample distance categories or 'bins'.
- The maximum inter-sample distance to be considered in variogram analysis (Reimann et al. (2008) recommend the actual maximum inter-sample distance × approximately 0.4).

- The function used to model the variogram: exponential, spherical, Gaussian, etc.
- The mathematical algorithm for fitting the function to the variogram, e.g. using least squares, or maximum likelihood, or even heuristically.
- Whether the fitting process should be weighted (and there are several options).
- Whether to fix key variogram parameters such as the nugget or sill (see Fig. 3.13).
- Whether to assume an underlying trend in the data.
- Whether the variation with distance is the same in all directions (i.e. whether an isotropic or anisotropic variogram model should be used)
- .. . and so on.

Variogram model fitting can, therefore, appear to be more like an art than a science (Bohling 2005). Alternative forms of interpolation (e.g. splines, inverse distance) to predict soil properties between sampling points are not universally recommended, however, and can result in unusual predictions depending on the mathematical interpolation method used.

3.4 Comparison of Sampling Strata

3.4.1 Comparing Mean or Median Values

The most convenient way of comparing between strata is to use some form of statistical means comparison (Fig. 3.14). The comparisons are *univariate*, in that mean values are compared one variable at a time. The first step of this analysis would be to assess some parameters describing the distribution of the variable of interest. We do this because standard statistical means comparison methods such as a t-test and analysis of variance (ANOVA) assume that the variable is normally distributed and that the variance is approximately equal in each stratum. We use a method such as the Shapiro-Wilk test to test for normality (against the null hypothesis that the distribution is not different from a normal distribution); it may be possible to transform variables to achieve normality (see Sect. 3.4.2). We also apply the Bartlett test or equivalent to test for heteroscedasticity (against the null hypothesis that the variance in each stratum is equal, i.e. homoscedastic). The conventional (parametric) statistical tests include the t-test (commonly implemented as Welch's t, for heteroscedastic variables) for two-level comparisons and the f-test as either standard ANOVA or Welch's f for heteroscedastic variables for comparisons with three or more levels. If the assumptions of normally distributed variables are not met, then non-parametric comparisons such as the Wilcoxon (for two-level comparisons) or Kruskal-Wallis (for multiple-level comparisons) tests can be used. The null hypothesis for all means comparison tests is that means in each stratum are equal, and the result of the tests is the probability that the null hypothesis is true for the population, given the values and variability that we have measured in our sample.

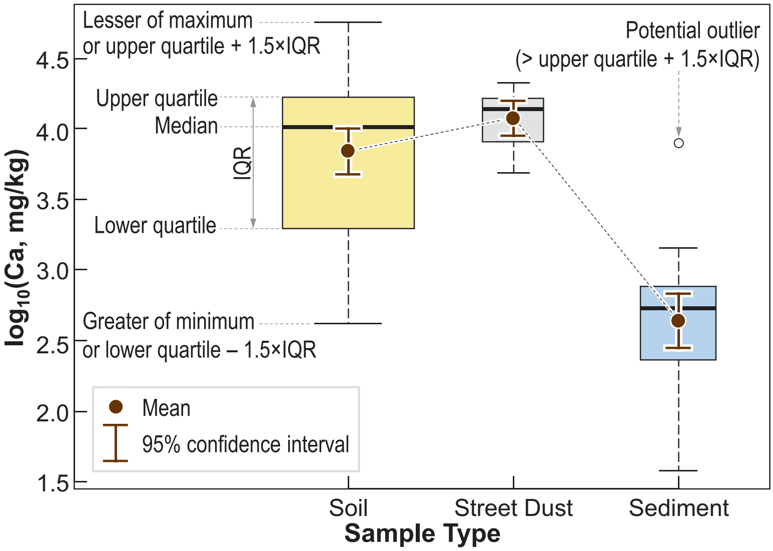


Fig. 3.14 Graphical comparisons of means and median concentrations of calcium in different sample types in an urban parkland, using an annotated and enhanced version of a standard 'Tukey' box plot. IQR = interquartile range (data and graphic by Andrew W. Rate)

3.4.2 Transforming Variables

In order to meet the assumptions of parametric statistical tests (e.g. t-tests, ANOVA/f-test), variables should be normally distributed. This is seldom the case for soil measurements (except sometimes soil pH), which commonly show positively skewed distributions. For continuous positively skewed variables, such as the concentration of a soil constituent, we tend to either use transform variables to logarithms (base 10 is most convenient and interpretable) or use a power function of the variable for transformation ($x_{\text{transformed}} = x^a$, where a is the power term). The power term which transforms a variable to have a distribution 'closest' to normal can be estimated from the Box-Cox algorithm, which is implemented in most statistical software. The power term a can be negative, which reverses the ordering of the variable (i.e. the greatest value will become the least and vice versa). In this case the ordering of the original variable can be preserved by including a factor of −1 in the power transform calculation ($x_{\text{transformed}} = -(x^a)$).

For different types of variables, particular transformations are required. For example, variables which are counts rather than continuous variables should not be transformed; instead alternative statistical models such as generalised linear models assuming a Poisson or negative binomial distribution should be used (O'Hara and Kotze 2010). Compositional variables (including concentrations or proportions of land surface coverage) are technically part of a fixed-sum closed set. For example, with data on percent land use over an urban area, all percentages add up to 100%! If uncorrected, fixed-sum closure can lead to very misleading conclusions, especially when relationships between variables are being investigated, as in correlation analyses or multivariate methods such as principal component analysis. Closed data require specialised transformations to remove closure, such as calculation of centred or additive log ratios (Reimann et al. 2008). The example in Fig. 3.15 shows the relationship between phosphorus (P) and iron (Fe) in soil/sediment materials in an acid sulphate environment. Without correcting for compositional closure, the P vs. Fe plot implies that P increases as Fe increases. Correcting for compositional closure, however, suggests the opposite, with P negatively related to Fe! In this case, if we had used conventional transformations, we might have come to a very wrong conclusion about the sediment properties affecting phosphorus.

The business of comparing means for environmental variables is possibly more complicated than we might have expected (as described above), but, to choose the correct method, the criteria are logical. The flow diagram in Fig. 3.16 shows how we can make the choice using three relatively simple questions: How many groups do we want to compare? Are our variables (transformed if necessary) normally distributed? And, does the variance of our variable depend on which group it is in? Once we have been guided in this way to the correct statistical test, we have two more questions that need to be asked. First, if we have more than two groups, which means are different from each other? We can answer this question rigorously using ‘pairwise comparison’ tests, as described below. Second, do we have a meaningful difference or just a statistical one? This is deciding whether we have a large or small ‘effect size’, and we discuss this below as well.

3.4.3 Pairwise Comparisons

The means comparison tests above will help us to make a decision about whether the mean value of some variable differs between strata. For sites with exactly two strata (where we would use a t-test or Wilcoxon test), the test tells us if the

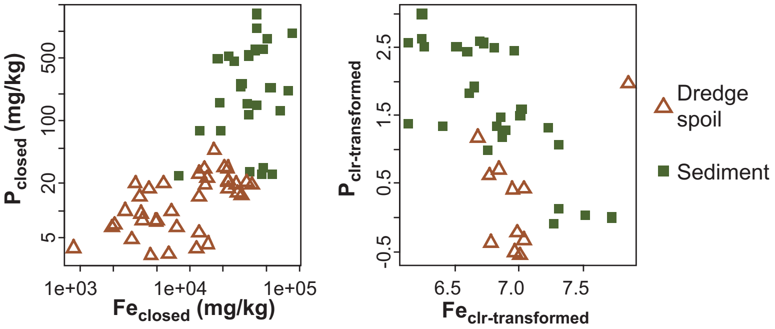


Fig. 3.15 Comparison of relationships between P and Fe for (**a**) compositionally closed concentrations showing a positive relationship and (**b**) concentration variables corrected for compositional closure using centred log ratios showing a negative relationship. Data from Xu et al. (2018); graphic by Andrew W. Rate

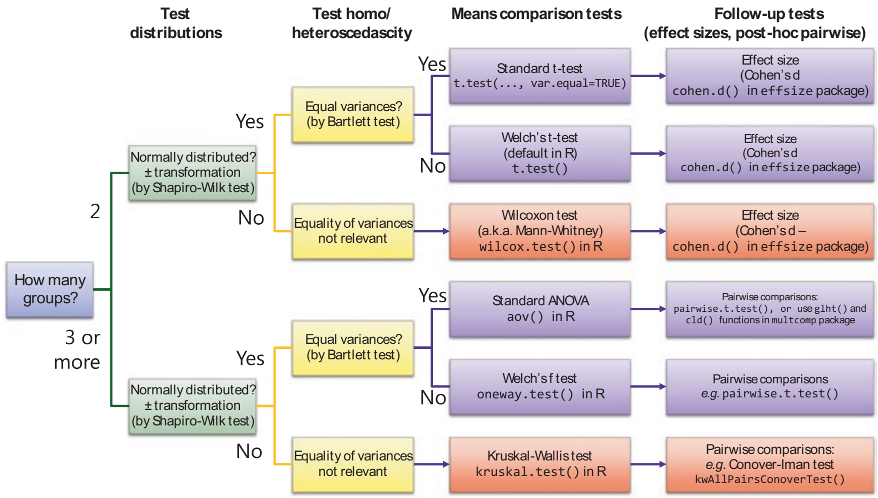

Test distributions
Test homo/ heteroscedascity
Means comparison tests
Follow-up tests (effect sizes, post-hoc pairwise)
How many groups?
2
3 or more
Normally distributed? ± transformation (by Shapiro-Wilk test)
Yes
No
Equal variances? (by Bartlett test)
Equality of variances not relevant
Yes
No
Standard t-test t.test(..., var.equal=TRUE)
Welch's t-test (default in R) t.test()
Wilcoxon test (a.k.a. Mann-Whitney) wilcox.test() in R
Effect size (Cohen's d cohen.d() in effsize package)
Effect size (Cohen's d cohen.d() in effsize package)
Effect size (Cohen's d – cohen.d() in effsize package)
Normally distributed? ± transformation (by Shapiro-Wilk test)
Yes
No
Equal variances? (by Bartlett test)
Equality of variances not relevant
Yes
No
Standard ANOVA aov() in R
Welch's f test oneway.test() in R
Kruskal-Wallis test kruskal.test() in R
Pairwise comparisons: pairwise.t.test(), or use glht() and cld() functions in multcomp package
Pairwise comparisons e.g. pairwise.t.test()
Pairwise comparisons: e.g. Conover-Iman test kwAllPairsConoverTest()

Fig. 3.16 Decision tree for choosing appropriate statistical tests for overall and pairwise comparisons of mean values of measurements between different sampling strata. Functions from the R statistical computing environment (R Core Team 2020) are shown for various tests in monospaced font (graphic by Andrew W. Rate)

difference in means is between all combinations of strata – since there is only one possible combination! In most sites where we study urban soils, we commonly have sites with three or more strata, and we want to compare mean values of our measurements between the strata. Ideally we would like to know which means are different from the others (not just if we can reject the null). So, if the f-test or Kruskal-Wallis test allows rejection of the statistical null hypothesis (i.e. equal means), we usually follow up with a pairwise test. For well-conditioned data where the assumptions of ANOVA are met, we can use the Tukey set of statistical analyses (least significant difference and rigorous pairwise p-values). For heteroscedastic variables this is not strictly allowed, but we can still apply something like a pairwise t-test with adjustment of p-values for multiple comparisons. Finally, with a non-parametric (e.g. Kruskal-Wallis) test of three or more means, we can use pairwise Wilcoxon tests with adjustment of p-values for multiple comparisons or a specialised pairwise comparison test such as Conover's test.

3.4.4 Effect Sizes for Comparing Means

We cannot rely on just rejection of the null hypothesis (of equal means), since it is mathematically possible for a statistically significant difference to exist when the practical difference is meaningless. In some cases an effect size statistic, of which the most common is Cohen's d (Eq. 3.1), can help us assess the magnitude of the difference between central values (means or medians). Cohen's d is a standardised measure of the difference between means for exactly two groups (e.g. strata), and its value is normally categorised as follows: d < 0.2 negligible; 0.2 < d < 0.5 small; 0.5 < d < 0.8 medium; d > 0.8 large. Most contemporary statistical software allows calculation of Cohen's d for binary (two group) comparisons; for multiple (pairwise) comparisons, some custom coding (e.g. in R) may be required. (Note that the size of the p-value for a t-test, ANOVA, or equivalent does not represent an effect size! We cannot assume we have a larger effect just because we have a smaller p-value.)

$$\text{Cohen's } d = \frac{\left(\text{mean}_{\text{group}\,1} - \text{mean}_{\text{group}\,2}\right)}{\text{Pooled standard deviations}} \tag{3.1}$$

3.5 Relationships between Variables

In the spatial context, we can use correlation or regression statistics to assess relationships between a soil variable and distance (e.g. distance from a potential or suspected source of contamination). We also investigate relationships between variables for other reasons, such as finding which observations do not follow the expected relationship, and we will look at an example of this in Chap. 6. For now, we will go through the basics of correct application of correlation and regression analyses.

3.5.1 Correlation Analysis

The most commonly used measure of correlation between two variables (bivariate) is Pearson's correlation coefficient, r, which can vary between −1 and 1, with an r value of zero meaning no correlation. The assumptions behind calculation of Pearson's r require that each variable is normally distributed; sometimes this can be achieved with an appropriate transformation (e.g. taking logarithms or a power function, bearing in mind the discussion in Sect. 3.4.2). Variables which are unable to be transformed to a normally distributed variable are unsuitable for Pearson's correlation analysis, but we can use a non-parametric method, Spearman's correlation, in such cases. Spearman's correlation is based on comparison of ranks within each ordered variable and is therefore independent of transformation. The p-value for correlation tests is for the null hypothesis of no relationship between the pair of variables, against the existence of a true correlation in the population from which the sample is taken.

Very often, it is useful for exploratory data analysis to generate a correlation matrix, which calculates a correlation coefficient (e.g. Pearson's or Spearman's) for all possible pairwise relationships between the variables selected. These are subject to the same requirements in terms of the distribution of variables as bivariate correlations, with the additional precaution that p-values should be adjusted upwards to account for the increased likelihood of type 1 errors (false positives) when multiple comparisons are made. Most statistical software will calculate these corrected p-values, for example, using Holm's method.

With any correlation analyses, it is essential to check the relationships graphically. It is easy to misinterpret r values if the data behave unexpectedly. For example, outliers may still exist in transformed variables, which have a large influence on the value of Pearson's r. The variables may show grouping or bimodality, so that the true relationships are masked by considering the data as whole or a strong relationship may exist which is non-linear as assumed in the Pearson correlation. Inspection of (appropriately transformed) bivariate plots can identify these types of issues, and most statistical software will allow plotting of scatterplot matrices to streamline this task.

3.5.2 Regression Analysis

If a relationship between variables exists, it should be possible to estimate, or predict, one variable from another. This prediction is the goal of regression models; in their simplest form of bivariate linear regression, they are conceptually similar to Pearson's correlation, but the focus is on the ability of the regression model (commonly a mathematical equation) to predict one variable, the 'dependent variable', from one or more 'predictor variables'. A thorough discussion of regression models would itself take a whole book, so we will not do that here! Instead we will look at a sequence of steps we can take to generate and assess different types of linear regression models, foreshadowing the example in Chap. 6, Fig. 6.10; the procedures in all of these steps should be available in any up-to-date statistical software. The aim of our regression model is to predict arsenic (As) concentration in soil, from the other soil measurements we have made. The dataset includes soil EC and pH, plus concentrations of numerous major and trace elements.

The general form of the multiple or simple linear regression models we will discuss is:

$$y = a + \sum_{i=1}^{n} \left(b_i x_i \right) + e \tag{3.2}$$

where y is the dependent variable to be predicted, n is the number of predictor variables (which can equal 1), a is the constant 'intercept' term, x_i are the predictor variables, b_i are the coefficients for each predictor, and e is the error term or 'residual'.

Initial Assumptions and Transformations For a valid linear regression model, the residuals need to be normally distributed. In practice we can increase the likelihood that this will be the case by transforming our variables to remove skewness, commonly with a $\log_{10}$ transformation. Since the goal is prediction, rather than analysis of the relationship itself, we can argue that issues like compositional closure can be ignored. In our example, we $\log_{10}$ transform all variables except soil pH which is used untransformed.

Choosing Predictors The next decision that needs to be made is what the predictor variable(s) should be. Realistically, we would normally try to predict a variable that is difficult, unreliable, or expensive to measure, since we would usually rather have an actual measurement than an estimate from prediction. The predictor variables, then, would logically be those which are more easily, reliably, and/or inexpensively measured. **More importantly**, we should try to choose predictors that make sense in the real world. For example, the concentrations of trace elements in soils are often closely related to one another due to similar geochemistry or common sources. In reality, though, it's unlikely that one trace element would have an effect on another in soils since the concentrations of both are too low. In soils, then, we tend to use 'bulk soil' properties, such as pH, clay content, organic carbon and other major element content, EC, redox potential, and so on, as predictors, since they fulfil both the 'easily measured' and 'realistic effect' criteria. For our example, we want to predict arsenic (As) concentrations, and our initial list of predictors is pH, EC, Al, Ca, Fe, K, Mn, Na, P, and S.

Collinearity of Predictors The predictors we select should not be linearly related to one another (collinear). The criteria we use to assess this are the Pearson correlations (which should be ≤0.8 between any pair of predictors) and variance inflation factors (VIF), which estimate how much greater the variance of a regression coefficient (the b_i values in Eq. 3.2) is, due to collinearity. There are various rules of thumb for selecting predictors on the basis of VIF: a value above 10 suggests that a predictor should be removed; 4 < VIF < 10 should be noted. In our example, the following pairs of predictor variables have Pearson's r greater than 0.8: Al-Fe, Fe-Mn, and Mn-P. The variance inflation factors are listed in Table 3.1. We will choose to remove Al and Mn from the 'maximal' regression model, but different variables could have been removed.

Refinement of Predictors Not all of the possible predictors that we select will have a significant influence on the value of our dependent variable. The output of statistical software (e.g. Table 3.2) usually has a p-value from a test of significance for each predictor (using the null hypothesis that the predictor has no effect on the dependent variable), as well as an analogous null hypothesis significance test for the model as a whole. Inspection of this output may imply that that some predictors have no

Table 3.1 Variance inflation factors (VIF) for a multiple regression model predicting As_{log} from pH, EC_{log}, Al_{log}, Ca_{log}, Fe_{log}, K_{log}, Mn_{log}, Na_{log}, P_{log}, and S_{log} (subscript $_{log}$ denotes $\log_{10}$ transformation)

Predictor	pH	EC_{log}	Al_{log}	Ca_{log}	Fe_{log}	K_{log}	Mn_{log}	Na_{log}	P_{log}	S_{log}
VIF	3.163	1.911	8.123	5.571	7.982	3.298	9.575	7.997	7.768	5.774

effect. To remove non-significant predictors, we use a stepwise regression algorithm, which systematically adds and removes predictors from a set of models, using an ‘information criterion’ to select the best subset of predictors which all contribute ‘information’ or predictive ability to the model. The stepwise algorithm can be configured to add predictors from a list to a basic model (forward selection) or to remove predictors from a maximal model or both. Ideally, different implementations of stepwise procedures, using the same data, should arrive at the same final answer.

We can see in Table 3.2 that the null hypothesis of no prediction ability is rejected at $p \leq 0.05$ for all predictors in the final model except S_{log} (this relates to the different selection criteria for predictors in the stepwise procedure). The model is good at predicting As concentration; the multiple R^2 (r-squared) value is 0.8155, so nearly 82% of the variance in $\log_{10}$As is explained by the four predictors. We can also reject (p-value = 2.6×10^{-16}, so $p \leq 0.05$) the null hypothesis of no prediction ability for the overall model. The VIF values are all close to 1, meaning negligible collinearity.

Model Checking Many of the assumptions for regression relate to the residuals, and we use a number of diagnostic tests and/or plots (Fig. 3.17) to assess these assumptions. First, the residuals have a median value close to zero (−0.0058, Table 3.2), and the mean residual value is 1.8×10^{-18}. By applying a Shapiro-Wilk test to the residuals from the model, we find that the null hypothesis (that the distribution is not different from a normal distribution) cannot be rejected, satisfying the assumption of normally distributed residuals.

Table 3.2 Summary of final regression model predicting As_{log} from initial predictors pH, EC_{log}, Al_{log}, Ca_{log}, Fe_{log}, K_{log}, Mn_{log}, Na_{log}, P_{log}, and S_{log} (subscript $_{log}$ denotes $\log_{10}$ transformation). Values explained in the text are in shaded cells with ***bold italic*** text. The same final set of predictors was obtained by either forward or backward stepwise selection of predictors in R (R Core Team 2020)

Residuals:	Min	1Q	Median	3Q	Max	
	−0.300155	−0.082543	***−0.00583***	0.049328	0.296157	
Coefficients:		Estimate	Std. error	t value	Pr(>|t|)	
	(intercept)	−1.08038	0.31638	−3.415	***0.001343***	**
	Fe_{log}	0.57608	0.05875	9.805	***7.58E-13***	***
	pH	−0.06438	0.01821	−3.535	***0.000944***	***
	EC_{log}	−0.16794	0.07197	−2.333	***0.024051***	*
	S_{log}	0.15322	0.07629	2.008	***0.050487***	.
Signif. Codes:	*** ≤ 0.001	** ≤ 0.01	* ≤ 0.05	. ≤ 0.1		
Statistics:	Residual standard error		0.1245 on 46 degrees of freedom			
	Multiple R-squared		***0.8155***	Adjusted R-squared		0.7994
	F-statistic	50.82 on 4 & 46 DF		p-value		***2.61E-16***
Predictor:	**Fe_{log}**	**pH**	**EC_{log}**	**S_{log}**		
VIF:	***1.374***	***1.192***	***1.028***	***1.562***		

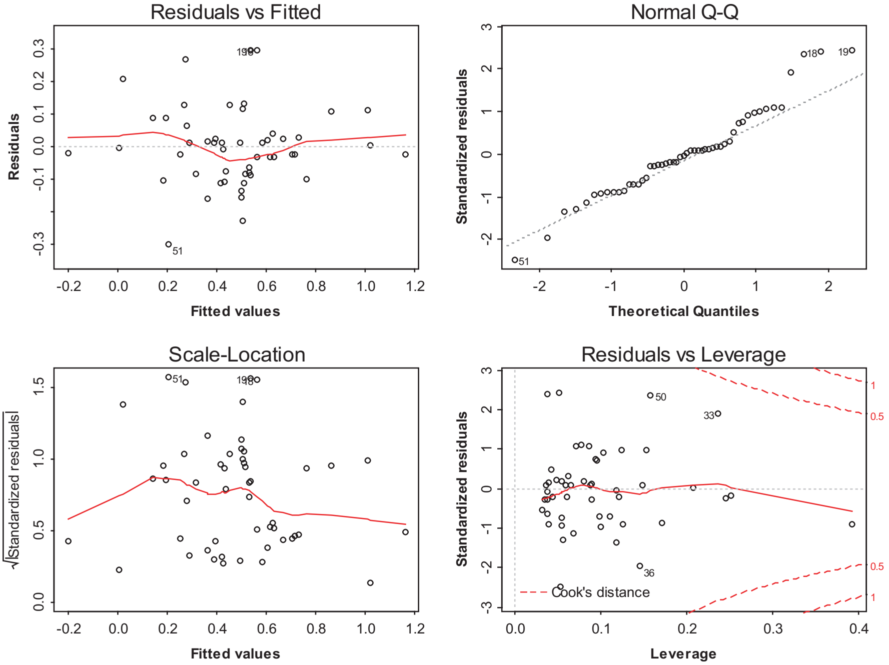


Fig. 3.17 Diagnostic plots for the final regression model predicting As_{log} from predictors pH, EC_{log}, Fe_{log}, and S_{log} (subscript log denotes log_{10} transformation)

The standard set of diagnostic plots (Fig. 3.17) allows us to assess, visually, some further regression assumptions. The residuals vs. fitted plot checks that the mean residual is close to zero and that there is no systematic trend in the residuals (this can be assessed separately by calculating residual autocorrelation; the autocorrelation coefficients should be close to zero). The normal Q-Q plot is a visual assessment of whether the residuals are normally distributed; the dotted straight line represents a perfect normal distribution with the same mean and standard deviation as the residuals, and the points lie approximately along this line, confirming the Shapiro-Wilk test result above. The scale-location plot assesses whether the residuals show homoscedasticity (i.e. the size of the residuals should be independent of the value of the dependent variable, measured or predicted). In our case there seems to be a 'bulge' of greater residuals in the middle of the plot, suggesting that this assumption may not be fulfilled for our model (again, we can test for this in more detail separately, e.g. using the Breusch-Pagan test (Hothorn et al. 2020)). Finally, the residuals vs. leverage plot is one way of testing if any individual observation has an unexpectedly large influence on the model parameters. Cook's distance is a measure of the change in regression parameters when a point is removed; ideally its value should be zero. There are a number of rules of thumb defining excessively large Cook's distance values, e.g. 4/n, where n is the number of observations (points).

The final regression model can be used in different ways. We can never use correlation or regression to make conclusions about whether one measurement causes another; …*correlation is not causation*. We could certainly use the soil pH, EC, and Fe and S contents, however, to predict As concentration with some accuracy. Actually, though, total As concentration is not so difficult to measure! We may choose to include regression models as part of more complex environmental simulation models where many parameters are required and we do not have access to data for all possible locations where prediction is required. One of the more powerful ways we can use regression models in urban environments is to make use of the deviations from the model – with well-chosen predictors, these can provide a good indication of truly unusual samples, and we look at an example of doing this in Chap. 6.

Of course, multiple regression is not the only variation on simple linear regression that we can make use of when studying urban soils. If we have different sampling strata (see Sect. 3.2.2 above), we can make use of our stratified sampling design in regression. We would not necessarily expect the same linear relationships between variables in different strata (which might, for instance, include both industrial land and undisturbed nature reserves). In this case we can use grouped linear regression (see Fig. 3.18 and Table 3.3), which effectively includes a separate intercept and coefficient(s) for each stratum within our data.

The general form of the grouped or simple linear regression models is similar to that for multiple regression:

$$y = \sum_{i=1}^{n} \left(a_i + b_i x\right) + e \tag{3.3}$$

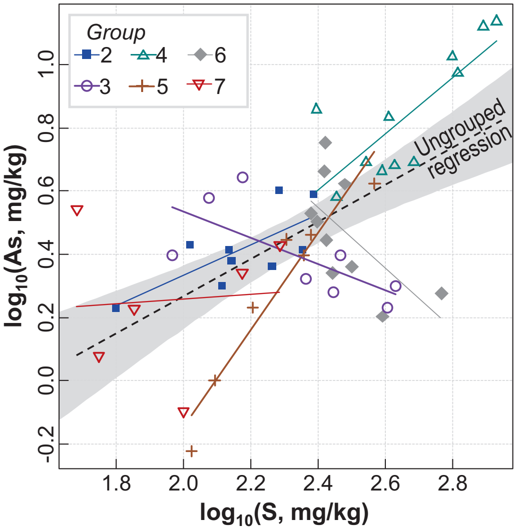


Fig. 3.18 Regressions predicting $\log_{10}$As from $\log_{10}$S, showing grouped (separate symbols, solid lines) and ungrouped (dashed line with shaded 95% confidence interval) regression models

Table 3.3 Summary and interpretation of ungrouped and grouped regression statistics

Ungrouped regression					
logAs = −0.905 + 0.587 · logS					
R^2 = 0.371, F statistic = 28.86 on 1 and 49 DF, p-value: 2.13 × 10^{-6}					
Grouped regression					
logAs = −0.621 + 0.478 · logS (group 2)					
logAs = 1.38 − 0.419 · logS (group 3)					
logAs = −1.54 + 0.894 · logS (group 4)					
logAs = 2.85 + 1.53 · logS (group 5)					
logAs = 2.85 − 0.959 · logS (group 6)					
logAs = 0.11 + 0.0742 · logS (group 7)					
R^2 = 0.802, F statistic = 14.3 on 11 and 39 DF, p-value = 1.72 × 10^{-10}					
Comparison of models					
Res.Df RSS	Df	sum of Sq	F	Pr(>F)	
Ungrouped	49	2.43273			
Grouped 39	0.76705	10	1.6657	8.469	3.928e-07

(P-values is ≤0.05 so the null hypothesis, that the more complex model makes no improvement in prediction, can be rejected.)

where the terminology is as for Eq. 3.3, except that now we have a single predictor x with different intercepts (a_i) and slopes (b_i) for each group of observations.

We should always check if the more complex model is actually better at prediction or whether it is simply ‘over parameterised’. We can compare linear regression models by using analysis of variance (Table 3.3) if they are nested, that is, a simpler model is a subset of a more complex model.

3.5.3 *Multivariate Analysis*

It is quite common to measure many variables in studies of urban soils. In the sections above, we have discussed how to analyse a dataset to interpret one variable at a time using univariate methods (although use of multiple linear regression does potentially use many variables to explain one other variable). Using various types of ordination analysis, we can use the information contained in multiple variables to create a reduced subset of variables containing nearly the same amount of information. Ordination methods are also referred to, for this reason, as ‘data reduction’ methods and are commonly used for multivariate analysis.

One of the earliest and most widely used ordination methods for exploration and dimension-reduction of multivariate data is principal components analysis (PCA; see the explanation in Box 3.2 and Fig. 3.19). Imagine a dataset with many samples (rows) and n continuous numeric variables (columns) which contain quantitative

Box 3.2 Principal Components Analysis

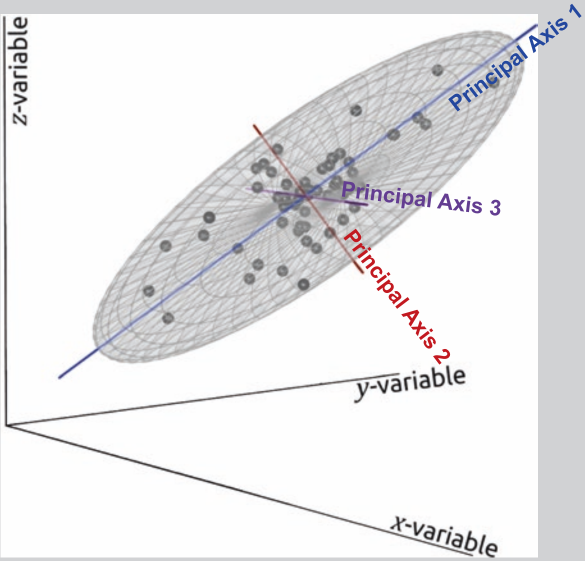


Fig. 3.19 Visualisation of principal components analysis for three variables/dimensions x, y, and z: we can conceptualise an ellipsoid encapsulating the 'cloud' of points (i.e. samples). The longest dimension (Principal Axis 1) of the ellipsoid, which accounts for the greatest proportion of multiple variance, is in the direction of its major axis and is a function of the variables x, y, and z. Principal Axis 2 must be orthogonal to Principal Axis 1 and is a different function of the variables x, y, and z which accounts for the next highest-possible proportion of multiple variance. Principal Axis 3 must be orthogonal to both Principal Axes 1 and 2; is a unique function of x, y, and z; and accounts for the remainder of multiple variance. For n variables/dimensions (n > 3), the analogy is an n-dimensional hyper-ellipsoid which has n orthogonal axes, which is very difficult to visualise

information about each sample such as concentrations, heights, velocities, etc. For these n variables/dimensions, the principal component calculation generates n new variables, or principal components, which are each a function of the set of all the original variables (so each principal component is defined by a weighting or coefficient for each of the original variables). We may choose to omit some variables from the analysis if they contain too many missing observations or if there is another valid reason to doubt their integrity. Since each principal component is selected to account for successively smaller proportions of the multiple variance, it is usually the first few principal components which explain most of the variance and therefore contain the most useful information. We conventionally visualise this in a 'scree plot' (Fig. 3.20a), a kind of bar graph showing the decrease in variance accounted for by each component (the 'eigenvalue').

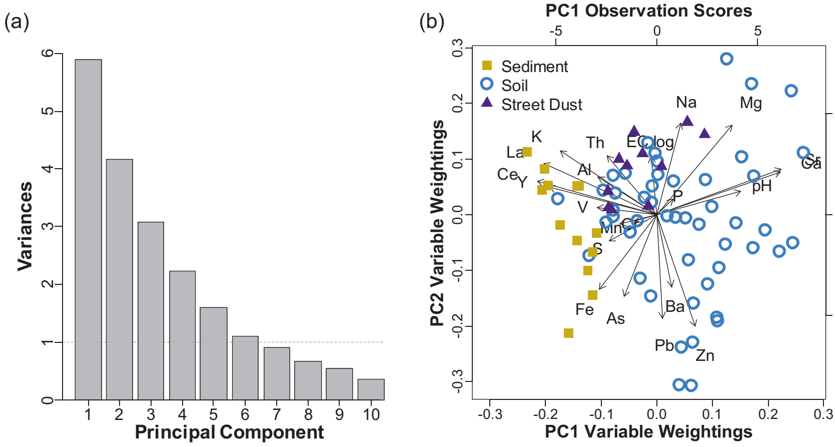


Fig. 3.20 (**a**) Scree plot of component variances and (**b**) biplot of the first two components, for principal components analysis on variables related to chemical properties of samples from an urban parkland. Concentration variables were transformed to centred log ratios to remove compositional closure prior to calculation of principal components on scaled, zero-centred variables (graphic by Andrew W. Rate)

As well as the component variances, the useful results of principal components analysis include the variable weightings or ‘rotations’ for each principal component. In addition, every individual observation (sample) is a multivariate point, the observation scores for all samples in each principal component based on the values of the variables used in PCA for that sample. It is conventional to plot both of these two types of output in a principal component ‘biplot’, as shown in Fig. 3.20b. Before discussing the biplot, we should note that the sign (positive or negative) of variable weightings and observation scores (i.e. the direction of the axes) is arbitrary and should not affect our interpretations.

The principal component biplot is useful because the variable weightings group together for variables (measurements) that are related to one another. For example, in the biplot in Fig. 3.20b, the variables are mainly concentrations of elements (which have been corrected for compositional closure before PCA using a log ratio transformation). These variables are shown as vectors (arrows) in the biplot of principal components PC1 and PC2, and elements which are geochemically related have vectors of similar length and/or direction. For example, the elements La, Ce, and Y are all geochemically similar rare-earth elements and plot closely together on the biplot, and the same is true for Ca and Sr which commonly co-occur in carbonate minerals. The other main information we obtain from principal components biplots is from the observation scores. These will plot at locations similar to their dominant variables: for example, in Fig. 3.20b, the sediment samples all plot towards the left of the biplot in the same direction as the K, La, Ce, Y, S, and Fe variable weighting vectors. This suggests that they are characterised by greater values of these variables (e.g. wetland sediments may contain higher concentrations of

Fe, S, and rare-earth elements due to formation of sulphides and Fe and S cycling (Morgan et al. 2012)).

We have used an example based on soil chemical data, but many other types of numerical data can be used in principal components analysis. These types of datasets could include soil physical data, composition of vegetation or microbial communities, and so on. For example, a dataset with variables which measure different plant species composition might show, in a PCA biplot, grouping of wetland plant species in riparian sampling strata, weedy species in disturbed urban land, native species in reserves, and so on. (Note that variables such as percent species compositions would comprise a fixed-sum closed set and would require a transformation to remove closure before rigorous multivariate data analysis such as PCA!) The provision of information of this type makes ordination methods such as principal components analysis powerful tools for exploratory data analysis. Different algorithms for ordination of multivariate data are based on different criteria than the maximisation of multiple variance used in PCA, for example, similarity or dissimilarity between samples. We will discuss different multivariate methods for analysis of data related to soil microbiology in Chap. 8.

3.6 Further Reading

Oliver MA, Webster R (2014) A tutorial guide to geostatistics: computing and modelling variograms and kriging. Catena 113:56–69. https://doi.org/10.1016/j.catena.2013.09.006

Reimann C, Filzmoser P, Garrett RG, Dutter R (2008) Statistical data analysis explained: applied environmental statistics with R. John Wiley & Sons, Chichester, England, 343 pp

3.7 Summary

- Cities affect soil variability on a subcontinental or regional scale, reflecting the concentration of human populations and resources in urban environments. Cities themselves contain variable soils on the scale of whole metropolitan areas, localities, sites, and individual soil profiles, and these are related to the age of human habitation and the types of activities conducted. Variation in soil properties with depth can inform us about site history or the extent of anthropogenic additions to soil.
- Sampling of soil needs to match the objective, that is, the type of information required. To capture regional- or metropolitan-scale soil variability, large systematic sampling exercises including hundreds or even thousands of samples are conducted. Studies of smaller spatial scales requires tens of samples at higher density, and depth variability is assessed with 2–30 vertical increments. Specific

approaches are required to detect discrete hotspots or sample pre-existing spatial strata.

- Analysis of spatial data begins with visual analysis in the form of maps with layer(s) of soil property data or scatterplots vs. distance along transects. More rigorous spatial data analysis techniques include variations of spatial autocorrelations or construction of variograms which allow spatial prediction using kriging.
- Other trends in spatial data, such as differences in soil parameters between strata, can be assessed with rigorously applied standard statistical techniques, both parametric and non-parametric, for comparison of central tendencies, assessing relationships, regression models, or multivariate ordination. For credible interpretation of urban soil data, care must be taken to ensure that the assumptions of each statistical method are met.

3.8 Questions

3.8.1 Checking your Understanding

1. What are the spatial scales of soil variability that we have considered? Are there any other scales that might be important in urban soil environments (your answer might differ depending on whether your focus is on ecosystem services, soil management, or soil research)?
2. How does the concept of an anthroposequence relate to the analysis (e.g. qualitatively or using quantitative measures like autocorrelation) of soil variability?
3. How many samples would you need to collect to have a 95% chance of detection of a circular contamination hotspot, 5 m in radius, over an area of 4 hectares? How far apart would the samples need to be?
4. List the advantages and disadvantages of the following four sampling designs: grid, stratified, random in grid, and completely randomised.
5. What is spatial autocorrelation measured with global Moran's I? What does the value of local Moran's I tell us about the relationships between samples which are close to one another?
6. Summarise the situations (i.e. properties of the variables and factors in a dataset) in which you would use the following mean/median comparison tests: Student's t, Wilcoxon, analysis of variance, and Kruskal-Wallis. When would you apply a pairwise comparison test?

3.8.2 Thinking about the Issues

7. Would it be reasonable to expect steep (sudden) gradients in soil properties between adjacent sampling strata or would soil properties show a more gradual change? Explain your answer, and try to think of an example of where the opposite to your initial answer might be true.
8. Figure 3.4 presents spatial data mapped as a continuous surface, whereas Fig. 3.6 presents similar data as point symbols containing the soil property information but without interpolation. Comment on the pros and cons of each approach.
9. How many reasons can you think of for using multivariate data analysis methods instead of multiple applications of conventional (uni- or bivariate) methods?

3.8.3 Using Your Creative Brain

10. Imagine that you are an expert witness, and a land developer has used a mean comparison to show that the average concentration (based on 50 soil samples) of polycyclic aromatic hydrocarbons (PAHs) on the land for the proposed development is not significantly different from 10 identical measurements on ‘background’ soil. They further argue that this means that there is no need for concern. Tell the hearing what is wrong with the land developer’s reasoning, and suggest what a more appropriate analysis of the data would be.

References

Anselin L (1995) Local indicators of spatial association—LISA. Geogr Anal 27:93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Bauer ME, Loffelholz BC, Wilson B (2008) Estimating and mapping impervious surface area by regression analysis of Landsat imagery. In: Weng Q (ed) Remote sensing of impervious surfaces. CRC Press / Taylor & Francis, Boca Raton, pp 3–20

Bohling G (2005) Introduction to Geostatistics and Variogram analysis, Kansas Geological Survey

Bu QW, Zhang ZH, Lu S, He FP (2009) Vertical distribution and environmental significance of PAHs in soil profiles in Beijing, China. Environ Geochem Health 31:119–131. https://doi.org/10.1007/s10653-008-9171-z

Bugdalski L, Lemke LD, McElmurry SP (2014) Spatial variation of soil lead in an urban community garden: implications for risk-based sampling. Risk Anal 34:17–27. https://doi.org/10.1111/risa.12053

Carr R, Zhang C, Moles N, Harder M (2008) Identification and mapping of heavy metal pollution in soils of a sports ground in Galway City, Ireland, using a portable XRF analyser and GIS. Environ Geochem Health 30:45–52. https://doi.org/10.1007/s10653-007-9106-0

Cicchella D, Giaccio L, Dinelli E, Albanese S, Lima A, Zuzolo D et al (2015) GEMAS: spatial distribution of chemical elements in agricultural and grazing land soil of Italy. J Geochem Explor 154:129–142. https://doi.org/10.1016/j.gexplo.2014.11.009

Coleman DC, Crossley DA Jr, Callaham MA (2017) Fundamentals of soil ecology. Elsevier Science & Technology, Saint Louis

Corrò E, Mozzi P (2017) Water matters. Geoarchaeology of the city of Adria and palaeohydrographic variations (Po Delta, Northern Italy). J Archaeol Sci Rep 15:482–491. https://doi.org/10.1016/j.jasrep.2016.08.001

Department of Environmental Protection (2001) Development of sampling and analysis programs. Contaminated Sites Management Series. Government of Western Australia, Perth, Western Australia, 69 pp.

Diawara DM, Litt JS, Unis D, Alfonso N, Martinez LA, Crock JG et al (2006) Arsenic, cadmium, lead, and mercury in surface soils, Pueblo, Colorado: implications for population health risk. Environ Geochem Health 28:297–315. https://doi.org/10.1007/s10653-005-9000-6

Fellows I (2019) OpenStreetMap: Access to open street map raster images, using the JMapViewer library by Jan Peter Stotz.. (R Package Version 0.3.4) http://CRAN.R-project.org/package=OpenStreetMap http://blog.fellstat.com/?cat=15. Accessed 20210616

Fletcher T, Duncan H, Poelsma P, Lloyd S (2004) Stormwater flow and quality, and the effectiveness of non-proprietary Stormwater treatment measures : a review and gap analysis. Technical Report 04/8, Melbourne, Australia

Gong Q, Deng J, Xiang Y, Wang Q, Yang L (2008) Calculating pollution indices by heavy metals in ecological geochemistry assessment and a case study in parks of Beijing. J China Univ Geosci 19:230–241. https://doi.org/10.1016/S1002-0705(08)60042-4

Gong M, Wu L, Bi X-Y, Ren L-M, Wang L, Ma Z-D, Li Z-G (2010) Assessing heavy-metal contamination and sources by GIS-based approach and multivariate analysis of urban–rural topsoils in Wuhan, Central China. Environ Geochem Health 32:59–72. https://doi.org/10.1007/s10653-009-9265-2

Grimm NB, Grove JG, Pickett STA, Redman CL (2000) Integrated approaches to long-term studies of urban ecological SystemsUrban ecological systems present multiple challenges to ecologists—pervasive human impact and extreme heterogeneity of cities, and the need to integrate social and ecological approaches, concepts, and theory. Bioscience 50:571–584. https://doi.org/10.1641/0006-3568(2000)050[0571:IATLTO]2.0.CO;2

Hothorn T, Zeileis A, Farebrother RW, Cummins C (2020) lmtest: Testing Linear Regression Models. R package version 0.9-38 https://CRAN.R-project.org/package=lmtest. Accessed 2021-06-03

Jim CY (1998) Soil characteristics and management in an urban park in Hong Kong. Environ Manag 22:683–695. https://doi.org/10.1007/s002679900139

Johnson CC, Ander EL (2008) Urban geochemical mapping studies: how and why we do them. Environ Geochem Health 30:511. https://doi.org/10.1007/s10653-008-9189-2

Kalogirou S (2019) Spatial autocorrelation, Vignette for R package 'lctools', https://cran.r-project.org/web/packages/lctools/vignettes/SpatialAutocorrelation.pdf. Accessed 20191204

Li X, Lee SL, Wong SC, Shi W, Thornton I (2004) The study of metal contamination in urban soils of Hong Kong using a GIS-based approach. Environ Pollut 129:113–124. https://doi.org/10.1016/j.envpol.2003.09.030

Lv J, Liu Y, Zhang Z, Dai J, Dai B, Zhu Y (2015) Identifying the origins and spatial distributions of heavy metals in soils of Ju country (Eastern China) using multivariate and geostatistical approach. J Soils Sediments 15:163–178. https://doi.org/10.1007/s11368-014-0937-x

Mann A, Reimann C, de Caritat P, Turner N, Birke M, Team GP (2015) Mobile Metal Ion® analysis of European agricultural soils: bioavailability, weathering, geogenic patterns and anthropogenic anomalies. Geochem Exploration Environ Analysis 15:99–112. https://doi.org/10.1144/geochem2014-279

Morgan B, Rate AW, Burton ED (2012) Trace element reactivity in FeS-rich eutrophic estuarine sediments: influence of formation environment and acid sulfate soil drainage. Sci Total Environ 438:463–476. https://doi.org/10.1016/j.scitotenv.2012.08.088

Nero BF, Anning AK (2018) Variations in soil characteristics among urban green spaces in Kumasi, Ghana. Environ Earth Sci 77:317. https://doi.org/10.1007/s12665-018-7441-3

Nikolaeva O, Rozanova M, Karpukhin M (2017) Distribution of traffic-related contaminants in urban topsoils across a highway in Moscow. J Soils Sediments 17:1045–1053. https://doi.org/10.1007/s11368-016-1587-y

Nriagu JO (1988) A silent epidemic of environmental metal poisoning? Environ Pollut 50:139–161. https://doi.org/10.1016/0269-7491(88)90189-3

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. Methods Ecol Evol 1:118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

Oliver MA, Webster R (2014) A tutorial guide to geostatistics: computing and modelling variograms and kriging. Catena 113:56–69. https://doi.org/10.1016/j.catena.2013.09.006

Pickett STA, Cadenasso ML, Grove JM, Nilon CH, Pouyat RV, Zipperer WC, Costanza R (2001) Urban ecological systems: linking terrestrial ecological, physical, and socioeconomic components of metropolitan areas. Annu Rev Ecol Syst 32:127–157. https://doi.org/10.1146/annurev.ecolsys.32.081501.114012

Pouyat RV, McDonnell MJ (1991) Heavy metal accumulations in forest soils along an urban- rural gradient in Southeastern New York, USA. Water Air Soil Pollut 57-58:797–807. https://doi.org/10.1007/BF00282943

Pouyat RV, Yesilonis ID, Russell-Anelli J, Neerchal NK (2007) Soil chemical and physical properties that differentiate urban land-use and cover types. Soil Sci Soc Am J 71:1010–1019. https://doi.org/10.2136/sssaj2006.0164

Pouyat RV, Yesilonis ID, Szlavecz K, Csuzdi C, Hornung E, Korsós Z et al (2008) Response of forest soil properties to urbanization gradients in three metropolitan areas. Landsc Ecol 23:1187–1203. https://doi.org/10.1007/s10980-008-9288-6

Powell RL, Roberts DA, Dennison PE, Hess LL (2007) Sub-pixel mapping of urban land cover using multiple endmember spectral mixture analysis: Manaus, Brazil. Remote Sens Environ 106:253–267. https://doi.org/10.1016/j.rse.2006.09.005

QGIS.org (2020) QGIS geographic information system. Open source geospatial foundation project https://qgis.org. Accessed 20210616

R Core Team (2020) R: A language and environment for statistical computing (Version 4.0.3 'Bunny-Wunnies Freak Out'). R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org http://www.R-project.org. Accessed 20210616

Rate AW (2018) Multielement geochemistry identifies the spatial pattern of soil and sediment contamination in an urban parkland, Western Australia. Sci Total Environ 627:1106–1120. https://doi.org/10.1016/j.scitotenv.2018.01.332

Reimann C, Filzmoser P, Garrett RG, Dutter R (2008) Statistical data analysis explained: applied environmental statistics with R. Wiley, Chichester, 343 pp

Ribeiro Jr. PJ, Diggle PJ, Schlather M, Bivand R, Ripley B (2020) geoR: Analysis of Geostatistical Data. {R package version 1.8.1}, http://CRAN.R-project.org/package=geoR. Accessed 20201215

Schaetzl RJ, Anderson S (2005) Soils: genesis and geomorphology. Cambridge University Press, New York

Smith DB (2004) Appendix 1: soil sampling protocols for geochemical landscapes pilot studies. U.S. Geological Survey, Denver

Taylor MP, Mackay AK, Hudson-Edwards KA, Holz E (2010) Soil Cd, Cu, Pb and Zn contaminants around Mount Isa city, Queensland, Australia: potential sources and risks to human health. Appl Geochem 25:841–855

Tobler W (2004) On the first law of geography: a reply. Ann Assoc Am Geogr 94:304–310

Tran BC, Teil M-J, Blanchard M, Alliot F, Chevreuil M (2015) Fate of phthalates and BPA in agricultural and non-agricultural soils of the Paris area (France). Environ Sci Pollut Res 22:11118–11126. https://doi.org/10.1007/s11356-015-4178-3

Turer D, Maynard JB, Sansalone JJ (2001) Heavy metal contamination in soils of urban highways: comparison between runoff and soil concentrations at Cincinnati, Ohio. Water Air Soil Pollut 132:293–314. https://doi.org/10.1023/a:1013290130089

USEPA, 2002. Guidance on choosing a sampling Design for Environmental Data Collection for use in developing a quality assurance project plan. EPA QA/G-5S, United States Environmental Protection Agency, Washington, DC

Webster R, Oliver MA (1993) How large a sample is needed to estimate the regional variogram adequately? Geo 1:155–166

White RE (2006) Principles and practice of soil science: the soil as a natural resource. Blackwell Science Ltd, Malden

Wu C (2004) Normalized spectral mixture analysis for monitoring urban composition using ETM+ imagery. Remote Sens Environ 93:480–492. https://doi.org/10.1016/j.rse.2004.08.003

Wu C, Murray AT (2003) Estimating impervious surface distribution by spectral mixture analysis. Remote Sens Environ 84:493–505. https://doi.org/10.1016/s0034-4257(02)00136-0

Xu N, Rate AW, Morgan B (2018) From source to sink: rare-earth elements trace the legacy of sulfuric dredge spoils on estuarine sediments. Sci Total Environ 637-638:1537–1549. https://doi.org/10.1016/j.scitotenv.2018.04.398

Zhang MK (2004) Phosphorus accumulation in soils along an urban-rural land use gradient in Hangzhou, Southeast China. Commun Soil Sci Plant Anal 35:819–833. https://doi.org/10.1081/CSS-120030360

Zhang C, Luo L, Xu W, Ledwith V (2008) Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci Total Environ 398:212–221. https://doi.org/10.1016/j.scitotenv.2008.03.011

Zhao D, Li F, Wang R, Yang Q, Ni H (2012) Effect of soil sealing on the microbial biomass, N transformation and related enzyme activities at various depths of soils in urban area of Beijing, China. J Soils Sediments 12:519–530. https://doi.org/10.1007/s11368-012-0472-6

Zhu W, Egitto BA, Yesilonis ID, Pouyat R (2017) Chapter 5: soil carbon and nitrogen cycling and ecosystem service in cities. In: Lal R, Stewart BA (eds) Urban Soils. Advances in Soil Science. CRC Press/Taylor & Francis Inc., Boca Raton, pp 121–135