

14 Spatial Epidemiology and Disease Mapping

Why Space Matters in Health Data

Disease risk, health service access, and health outcomes are rarely distributed uniformly across space; they cluster according to environmental suitability (vector habitat, rainfall, temperature), health-system access (distance to the nearest facility, road network quality), and socioeconomic patterning that is itself spatially structured. Ignoring spatial structure in health data analysis is not a neutral simplification — it typically means treating geographically neighbouring observations (adjacent LGAs, nearby facilities) as independent when in fact they share unmeasured environmental and social drivers, understating uncertainty in exactly the way that ignoring survey clustering does (Chapter 7).

Choropleth Mapping and Its Pitfalls

A choropleth map shades administrative units (states, LGAs) by the value of an indicator (leishmaniasis case counts, malaria TPR, vaccination coverage) and is the most common and most easily misused tool in spatial health communication. Three pitfalls recur: using raw counts rather than rates, which simply reproduces the population-density map rather than showing disease risk; choosing a colour classification scheme (equal interval, quantile, natural breaks/Jenks) that can visually exaggerate or obscure real differences depending on the choice; and, most fundamentally,

the modifiable areal unit problem — the pattern visible in a map can change substantially depending on the arbitrary administrative boundaries chosen for aggregation (LGA versus ward versus state), a limitation that must be acknowledged whenever a choropleth map is presented as evidence of a spatial pattern.

R example

```
# R: choropleth map of leishmaniasis case rates by state, with attribution stamp
library(sf); library(ggplot2); library(dplyr)

nigeria_states <- st_read("nigeria_state_boundaries.shp")
map_data <- nigeria_states %>%
  left_join(nli_state_summary, by = c("state_name" = "state")) %>%
  mutate(rate_per_100k = (cases / population) * 100000)

ggplot(map_data) +
  geom_sf(aes(fill = rate_per_100k), colour = "white", linewidth = 0.1) +
  scale_fill_viridis_c(name = "Cases per\n100,000", option = "magma") +
  labs(title = "Leishmaniasis Case Rate by State, Nigeria",
       caption = paste("Source: Nili Platform baseline data. Generated:", Sys.Date())) +
  theme_minimal()
```

Spatial Autocorrelation and Disease Clustering

Moran's I and related statistics formally test whether an indicator's values are more spatially clustered than would be expected under complete spatial randomness — a positive, statistically significant Moran's I for malaria TPR would indicate that high-TPR LGAs tend to be surrounded by other high-TPR LGAs more often than chance would predict, consistent with underlying shared environmental drivers of transmission. Local indicators of spatial association (LISA) extend this to identify specific hotspots and coldspots, which is directly actionable for targeting supplementary interventions (indoor residual spraying, active case-finding campaigns) to areas of unusually high, spatially clustered risk.

Spatial Regression and Disease Mapping Models

When spatial autocorrelation is present in the residuals of an ordinary regression model (Chapter 8), standard errors and, in

some cases, coefficient estimates are unreliable, and spatial regression extensions — the spatial lag model (where a location's outcome depends on neighbouring locations' outcomes) and the spatial error model (where unmeasured spatially correlated factors affect the error term) — correct for this. Bayesian disease mapping models, particularly the Besag-York-Mollié (BYM) model widely used in small-area disease mapping, borrow statistical strength across neighbouring areas to produce smoothed, more stable risk estimates for small administrative units (such as LGAs with low population or case counts) where a raw rate would be extremely noisy — directly analogous in spirit to the multiple-imputation and hierarchical-modelling logic introduced earlier in this book, in that both borrow information to stabilise estimates that would otherwise be unreliable in isolation.

Worked Example

Calculating and Interpreting Moran's I

Suppose malaria TPR values across 20 neighbouring LGAs are tested for spatial autocorrelation using a queen-contiguity neighbour matrix (LGAs sharing a border are considered neighbours). The observed Moran's I is 0.42, and a permutation test (999 random reshuffles of the TPR values across the same spatial arrangement) yields a p-value of 0.001, since none of the 999 random permutations produced an I as extreme as 0.42. This provides strong evidence that malaria TPR is spatially clustered — high-TPR LGAs are non-randomly located near other high-TPR LGAs — which supports the hypothesis of shared environmental or vector-ecology drivers operating at a scale larger than a single LGA, and argues for considering multi-LGA (rather than purely LGA-by-LGA) intervention planning in the identified hotspot zone.

Figure 14.1 — Illustrative Spatial Clustering of TPR Across LGAs
(Moran's $I \approx 0.42$, $p = 0.001$ — Sec. 14.5)

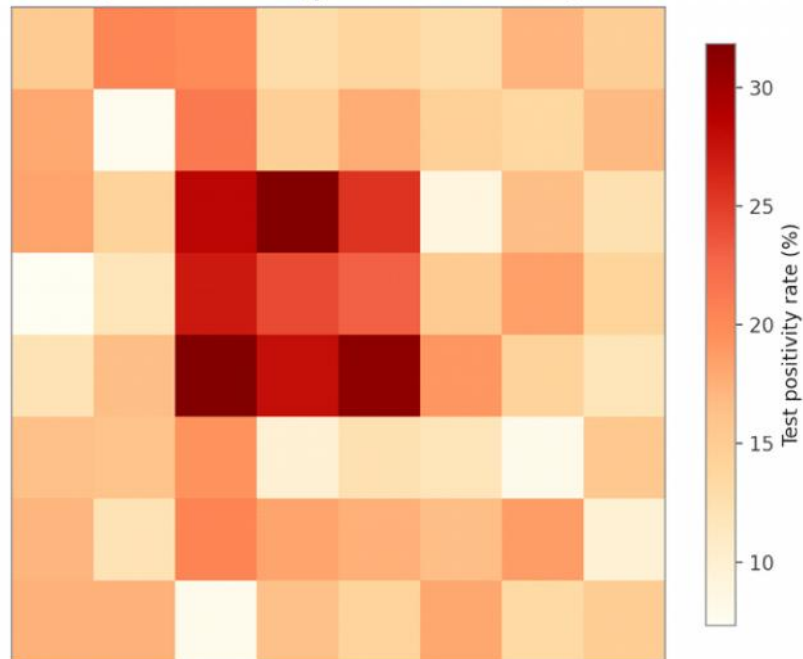


Figure 14.1 — Illustrative spatial clustering of TPR across LGAs
(Section 14.5).

PYTHON example

```
# Python: computing Moran's I with the esda/libpysal ecosystem
import libpysal
from esda.moran import Moran

w = libpysal.weights.Queen.from_dataframe(lga_gdf)    # queen-contiguity neighbours
w.transform = 'r'

moran = Moran(lga_gdf["tpr"].values, w, permutations=999)
print(f"Moran's I = {moran.I:.3f}, p-value = {moran.p_sim:.3f}")
```

Worked Example

A Simplified BYM-Style Smoothing Comparison

Consider three small LGAs with populations of 8,000, 12,000, and 150,000 and observed leishmaniasis case counts of 2, 3, and 18 respectively, giving raw rates of 25.0, 25.0, and 12.0 per

100,000 — the two small LGAs appear to have double the risk of the large one, but with so few cases, this could easily be chance. A simple empirical Bayes smoothing approach (a simplified relative of the full BYM model) shrinks each LGA's rate toward the overall regional average (here, roughly 13.5 per 100,000) in proportion to its uncertainty: the small LGAs' rates are pulled substantially toward 13.5 (smoothed estimates of approximately 16.8 and 17.5), while the large, more stable LGA's rate barely moves (smoothed to approximately 12.3), illustrating the core logic of small-area disease mapping — that raw rates from small populations are unreliable and should be stabilised by borrowing strength from the wider area, not taken at face value.

LGA	Population	Cases	Raw rate/100k	Smoothed rate/100k
A (small)	8,000	2	25.0	16.8
B (small)	12,000	3	25.0	17.5
C (large)	150,000	18	12.0	12.3

Geocoding and Common Spatial Data Pitfalls

Before any spatial analysis in this chapter can proceed, case or facility records must be correctly geocoded — matched to a geographic coordinate or administrative unit. Common pitfalls include facility name mismatches between a health facility registry and a case dataset (the same fuzzy-matching problem introduced in Chapter 4's Section 4.7, now applied to spatial linkage), coordinates recorded in the wrong hemisphere sign convention (a straightforward but surprisingly common error that places a Nigerian facility incorrectly south of the equator or west of the prime meridian), and administrative boundary files that are out of date relative to recent LGA or ward boundary changes — any of which can silently corrupt a spatial analysis if not checked before mapping.

Extended Case Study

A Hotspot That Moved When the Map Was Redrawn

A striking and pedagogically important illustration of the modifiable areal unit problem introduced in Section 14.2: a spatial analysis of childhood stunting prevalence, mapped at the LGA level, identifies a single LGA as a clear, isolated hotspot with a prevalence more than double its immediate neighbours. When the same underlying household-level survey data is instead aggregated to the finer ward level, the apparent hotspot dissolves into a more nuanced pattern: only two of the LGA's eleven wards actually show elevated stunting prevalence, while the remaining nine wards have prevalence closely in line with the surrounding LGAs — meaning the LGA-level aggregation had averaged two genuinely high-burden wards together with nine broadly typical wards, producing an LGA-level figure that looked like a uniform hotspot but that in fact concealed substantial within-LGA heterogeneity.

This has direct and significant practical consequences: an intervention planned on the basis of the LGA-level map alone would likely allocate supplementary nutrition resources uniformly across the entire LGA (including the nine wards that do not actually need them, at the expense of resources that could have been more intensively targeted at the two wards that do), whereas the ward-level analysis supports a much more efficient, precisely targeted intervention design. The broader lesson, consistent with this chapter's repeated caution about the modifiable areal unit problem, is that the finest administrative or geographic resolution at which reliable data can be obtained should generally be preferred for intervention targeting purposes, with coarser-level maps reserved mainly for high-level programme communication and overview, precisely because coarser aggregation can systematically conceal exactly the kind of actionable within-area heterogeneity that this case study demonstrates.

Edge Effects at Administrative and National Boundaries

Spatial analyses conducted strictly within Nigeria's national boundary, or within a single state's boundary, are subject to an edge effect: an LGA or state located directly on an international or inter-state border may show an apparently isolated, unexplained spatial pattern simply because the analysis has no visibility into disease dynamics or population movement occurring just across that boundary, artificially truncating what may in reality be a single, continuous cross-border transmission zone. This is a genuine practical concern for Nigeria's northern border states with respect to diseases such as Lassa fever and cholera, where transmission dynamics plausibly do not respect the administrative boundary at all, and it argues for actively seeking cross-border surveillance data-sharing arrangements wherever feasible, or at minimum for explicitly noting border-adjacent units as having incomplete spatial context in any Moran's I or hotspot analysis (Section 14.3), rather than treating the national or state boundary as if it represented a genuine epidemiological barrier.

Remote Sensing and Environmental Covariates in Disease Mapping

Beyond the administrative case-count data emphasised throughout this chapter, satellite-derived environmental covariates — rainfall estimates, land surface temperature, vegetation indices (which can proxy for standing water and vector breeding habitat suitability), and nighttime lights data (often used as a proxy for population density and urbanisation where census data is outdated) — are increasingly available at fine spatial resolution, free of charge, from sources such as NASA and the European Space Agency's satellite programmes, and can substantially enrich a spatial disease model by providing genuinely independent, environmentally-grounded covariates rather than relying solely on case counts and population denominators. Incorporating these

covariates into the spatial regression models of Section 14.4 — for example, testing whether a vegetation index proxy for vector habitat suitability predicts leishmaniasis case rates independent of population density — can both improve a model’s predictive accuracy for under-surveyed areas and provide genuine biological insight into transmission drivers, though it requires additional technical skill in remote sensing data processing (typically using specialised raster-data libraries) beyond the vector/tabular spatial data handling emphasised in this chapter’s core examples.

Kriging and Geostatistical Interpolation for Point-Referenced Data

Where Sections 14.2-14.6 dealt with data already aggregated to administrative units (states, LGAs), some Nigerian health data — environmental covariates from weather stations, or point-referenced prevalence survey cluster results — arrive instead as measurements at specific geographic points that must be interpolated to produce a continuous surface across unmeasured locations. Kriging, a geostatistical interpolation method, uses the spatial autocorrelation structure of the observed points (formalised through a variogram, describing how similarity between points decays with distance) to produce not only an interpolated value at any unmeasured location but also an explicit measure of the interpolation’s uncertainty, which tends to be far more statistically principled than simpler distance-weighted interpolation methods that provide no accompanying uncertainty estimate at all.

Spatiotemporal Modelling

Combining Chapters 10 and 14

Real disease surveillance data is neither purely spatial nor purely temporal but spatiotemporal — case counts vary across both LGAs and months simultaneously, and the two dimensions interact (a spatial hotspot in one season may shift location in the next).

Spatiotemporal models extend the spatial smoothing methods of Section 14.4 (such as BYM) to include a time dimension, allowing both spatial and temporal ‘borrowing of strength’ simultaneously — a small LGA’s estimate for a given month benefits both from its neighbouring LGAs’ data in that same month and from its own data in neighbouring months — producing considerably more stable estimates for small-area, high-frequency surveillance data than either a purely spatial or purely temporal smoothing approach applied in isolation, at the cost of a substantially more complex model requiring specialised software (such as R’s ‘INLA’ package) to fit efficiently.

A More Robust Spatial Analysis Function

The Moran’s I calculation in Section 14.5 assumes a correctly-formed neighbour matrix is already available; a more defensive version checks for the specific, common failure mode of an island polygon (an administrative unit with no detected neighbours, which silently breaks a queen-contiguity weights matrix) before running the test, since an undetected island produces a misleading result rather than a clear error, exactly the kind of silent spatial-data problem flagged more generally in Section 14.7.

PYTHON example

```
# Python: a Moran's I function that checks for islands before proceeding
import libpysal
from esda.moran import Moran

def morans_i_with_island_check(gdf, value_col):
    w = libpysal.weights.Queen.from_dataframe(gdf)
    islands = w.islands
    if islands:
        raise ValueError(f"{len(islands)} polygon(s) have no detected neighbours "
                        f"(indices: {islands}) – fix the geometry before testing")
    w.transform = "r"
    moran = Moran(gdf[value_col].values, w, permutations=999)
    return {"I": moran.I, "p_value": moran.p_sim}

result = morans_i_with_island_check(lga_gdf, "tpr")
```

Chapter Summary

- Spatial structure in health data reflects shared environmental and socioeconomic drivers across neighbouring areas, and ignoring it understates uncertainty.
- Choropleth maps must use rates (not raw counts), a deliberately chosen classification scheme, and an acknowledgement of the modifiable areal unit problem.
- Moran's I and LISA statistics formally test and localise spatial clustering, informing targeted intervention; a permutation test provides a defensible p-value for the observed clustering statistic.
- Spatial and Bayesian disease-mapping models (spatial lag/error models, BYM) correct for spatial autocorrelation and stabilise small-area estimates by borrowing strength from neighbouring areas.
- Small-area rates computed from low case counts and small populations should be smoothed (shrunk toward a wider regional average) rather than interpreted at face value.

Mini-Project

Mini-Project 14: A Spatial Analysis and Smoothing Report

Produce a spatial analysis of a disease indicator across a set of administrative units, including a map, a spatial autocorrelation test, and a small-area smoothing comparison.

Step 1. Obtain or simulate case counts and population denominators for at least 15 administrative units (LGAs or wards) with a neighbour structure (or simulate approximate adjacency).

Step 2. Produce a choropleth map of the rate (not raw count), choosing and justifying a classification scheme.

Step 3. Compute Moran's I (or describe the calculation by hand for a small example) and report whether the pattern is significantly spatially clustered.

Step 4. Identify at least two units with very small populations and compute both their raw and empirical-Bayes-smoothed rates, following the logic of Section 14.6.

Step 5. Write a short paragraph advising which units should be prioritised for intervention, explicitly distinguishing units that are 'genuinely high risk' from those that merely have a small, noisy denominator.

Deliverable: *One choropleth map, a Moran's I result with interpretation, and a smoothing comparison table with a written recommendation.*

End-of-Chapter Exercises

1. Explain why a choropleth map of raw leishmaniasis case counts by state could be misleading, and describe how you would correct it.
2. What does a positive, statistically significant Moran's I for malaria TPR across LGAs suggest about the underlying drivers of transmission?
3. Why might a BYM (Bayesian small-area) model produce a more reliable risk estimate for a low-population LGA than the raw observed rate?
4. Using the worked example in Section 14.5, explain in your own words what the permutation test p-value of 0.001 is actually testing (i.e., what null hypothesis is being compared against).
5. In Section 14.6, explain why LGA C's raw and smoothed rates are so close together, while LGA A and B's rates change substantially under smoothing.
6. Using the table in Section 14.6, if LGAA's population were instead 80,000 (ten times larger) with the same 2 cases, would you expect its smoothed rate to move further from or closer to its raw rate compared to the original scenario? Explain why.
7. A journalist wants to publish a map showing 'the most dangerous LGA for leishmaniasis' based on raw case rates alone. Write a short response explaining the risk of doing this without smoothing, using a concrete hypothetical example.
8. Using Section 14.7, identify two ways a facility name mismatch between a case dataset and a facility registry could cause a spatial analysis to silently produce incorrect results.

9. Why is a wrong hemisphere sign convention (e.g., a positive instead of negative latitude) a particularly dangerous kind of geocoding error to miss?
10. How would you check whether an administrative boundary shapefile you are using is out of date relative to recent LGA changes?
11. In the Section 14.8 case study, explain why the ward-level map revealed information the LGA-level map had concealed.
12. Why would an intervention based only on the LGA-level map have been less efficient than one based on the ward-level map?
13. At what point would going to an even finer geographic resolution (e.g., individual settlements) stop being useful, and why?
14. Propose one practical constraint that might prevent a health programme from always using the finest available geographic resolution for intervention targeting.
15. How would you decide, in general, which geographic resolution is 'fine enough' for a specific health indicator and intervention decision?
16. Using Section 14.9, explain why an LGA directly on Nigeria's northern border might show an apparently unexplained spatial pattern in a Lassa fever analysis.
17. Propose one practical step a state or national programme could take to reduce the edge effect problem described in Section 14.9.
18. Why might treating a state boundary as an epidemiological barrier be especially problematic for a disease like cholera, compared to a disease with very limited geographic mobility?

19. How would you flag a border-adjacent LGA's spatial analysis result as having 'incomplete spatial context,' in a technical report?
20. Explain why cross-border data-sharing arrangements, while logistically difficult, would directly address the root cause of the edge effect problem.
21. Using Section 14.10, propose one specific way a vegetation index could serve as a proxy for vector breeding habitat suitability in a leishmaniasis spatial model.
22. Why might nighttime lights data be a useful substitute for outdated census population figures in a rapidly urbanising area?
23. What additional technical skill does incorporating remote sensing covariates require, beyond the vector/tabular spatial methods emphasised elsewhere in Chapter 14?
24. Explain one potential benefit of incorporating environmental covariates for an under-surveyed area with limited case-count data.
25. Propose a specific research question that could be tested by combining case-count data with a satellite-derived rainfall covariate.
26. Summarise, in your own words, why this chapter treats the modifiable areal unit problem as a fundamental, not merely technical, limitation of spatial analysis.
27. Explain the common thread linking the edge-effects discussion (Section 14.9) and the moving-hotspot case study (Section 14.8).
28. Why does this chapter recommend incorporating remote sensing and environmental covariates (Section 14.10) rather than relying on case counts alone?

29. Propose a spatial analysis from your own field where edge effects at an administrative or national boundary (Section 14.9) could distort a naive interpretation.
30. If you had to explain to a journalist why ‘the most dangerous LGA’ can change depending on how the map is drawn, what would you say?
31. Using Section 14.12, explain why a spatiotemporal model can produce more stable small-area estimates than either a purely spatial or purely temporal model alone.
32. Describe a Nigerian surveillance dataset where a spatial hotspot might plausibly shift location from one season to the next.
33. Why does fitting a spatiotemporal model typically require more specialised software than fitting a purely spatial model?
34. What is meant by ‘borrowing strength’ in both the spatial and temporal dimensions simultaneously?
35. How does the added complexity of a spatiotemporal model relate to the sensitivity-analysis discipline discussed in Section 13.12?

Part 5

Big Data, Ethics, and Practice

