Field Guide of Relational and Contextual Epidemiology

Jon Zelner

2023-09-08

Table of contents

1 Inspirations and Influences

2 Inspirations and Influences

What - if any - kind of book needs to be written about relational and contextual epidemiology?

2.1 Why relational and contextual epidemiology?

There is a hole in the epidemiological literature that limits our ability to come to grips with the *relational* aspects of health and illness. For the purposes of this book, relationships are sources of non-independence. These could be *social* influences, e.g. from our friends and family. But they could also be *spatial* in nature, e.g. a result of some physically proximate environmental exposure. Often, the relationships we care about are *temporal* in nature, as is evident in the lifecourse perspective in which early-life exposures are understood to impact later-life outcomes.

There are many good books and papers out there about these topics as well as some of their conjunctions with each other (e.g. spatiotemporal analysis). But there is less written about their overlaps and commonalities, why knowing one teaches you so much about the others, and what having access to this set of tools might mean for a working epidemiologist.

In many ways, my goal in writing this is selfish, to satisfy a personal curiosity I suspect is of some greater import to at least a few people. Specifically: What does it mean to do *contextual* or *relational* epidemiology? Are these meaningless terms that just appeal to the social and natural science centers of my brain, or does this tap into something more meaningful?

More than methods

Part of my inspiration is a reaction to the technocratic impulse and imperative in modern epidemiology. In particular, I find myself a bit paralyzed by the worry about what happens when we operate under the assumption that escalating methodological complexity is an imperative and that the road out of socio-epidemiological problems is paved with technological solutions.

But also, methods...

On the other side, simply put, I love the methodological tools of spatial epidemiology, Bayesian hierarchical analysis, and systems modeling. I have learned more than I ever could have hoped through learning, tinkering with, and applying these tools to problems in the real world and in my own head. But for me, the large majority of lessons learned have been from their conceptual isomorphisms (or conflicts) with the world as it appears to us through qualitative and quantitative data, rather than in the exact values of their parameter values and quantitative predictions.

For me, this book is about resolving this cognitive dissonance while providing useful 'how-to' pointers along the way. I hope to articulate the affirmative case for a systems-based, contextually-sensitive, justice-oriented, morally and ethically opinionated, and theoretically driven approach to epidemiology. Along the way, I hope to show why the tools of such an approach are necessarily heterogeneous in nature and require us to accept uncertaintities quantitative and epistemic.

2.2 Influences

There are any number of books and papers out there that have articulated a similar perspective on the tools of quantitative analysis. The following have been particularly important for my own thinking and their influences will be felt throughout this work:

- Statistical Rethinking (1)
- The Ecological Detective (2)
- ARM/Regression and Other Stories (3)

What makes these works so useful, strong, and enduring is the way that they articulate a coherent, opinionated perspective on the meaning and use of a set of methodological tools. On top of that, they are engaging and fun to read - the sort of thing you return to over time not just to get specific methodological tools, but to be exposed to their perspective.

References

Part I

Introduction

3 An invitation

One theme that comes up consistently when I'm teaching and talking to students and others about topics in relational epidemiology - particularly through my courses on spatial analysis - is that it has an aura of inaccessibility.

This makes a certain kind of sense: Making maps and measuring and modeling spatial relationships might seem like it is outside of the classic analytic toolkit of working epidemiologists, not to mention other fields in public health, medicine, and the social sciences. In fact, this was my take on it as well, until I realized space has always been at the heart of my research, even when I didn't realize it!

The diagram in Figure 3.1 reflects my interpretation of the way someone coming to this area of research and practice for the first time might see it:

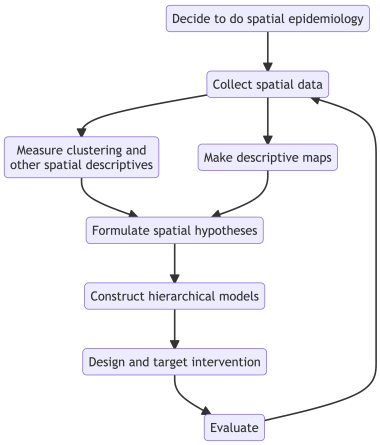


Figure 3.1: A textbook version of the spatial epidemiology workflow, without all the twists and turns that characterize actual research.

In this view, spatial epidemiology or analysis is taken to be a set of relatively fixed and well-described ideas and procedures connected to a highly technical set of methods. When we come from this perspective, starting a spatial project requires us to embark with a backpack that is already filled with specialized spatial tools.

3.1 You are (almost certainly) a relational analyst already

My goal in this short essay is to chip away a bit at the idea of spatial/relational epidemiology as something separate and apart from mainstream epidemiology and public health. Instead, I argue that these are better understood as a loose wrapper around a core set of ideas and tools that are part of the working arsenal of most professionals, students, and researchers in public health.

To explain what I mean, and why I think it's important, I'm going to subject you to a bit of autobiography about my own intellectual and professional trajectory.

3.2 Maps: Gateway or destination?

I began working as an epidemiologist or epidemology-adjacent type as a PhD student at the University of Michigan some time around 2006. As part of my dissertation research, I worked with Joe Eisenberg on an analysis ofthe role of social networks as sources of risk and protection against diarrheal disease and other infections in a group of villages in an area of rural Ecuador:

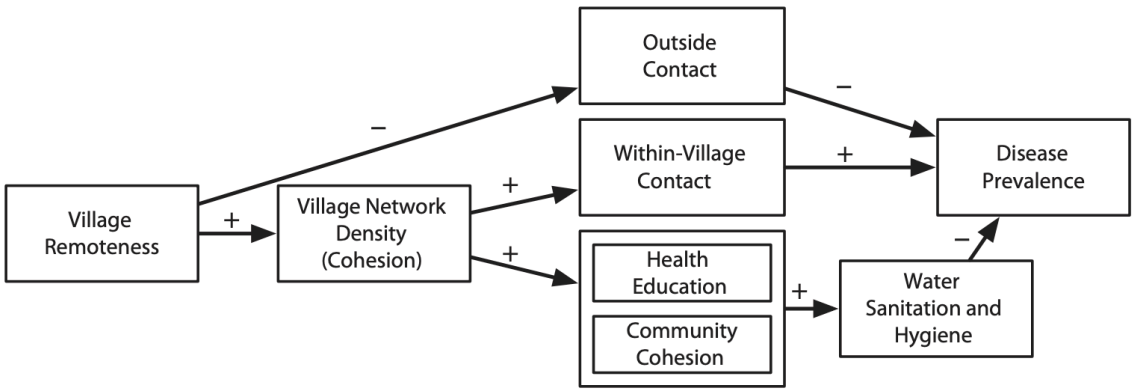


Figure 3.2: A figure from my dissertation research representing hypothesized relationships between village context (represented by inaccessibility or 'remoteness') and variation in disease outcomes within and between villages. (From (4))

Looking back, this was indisputably a spatial analysis: We were interested in how local social contexts impacted variation in health outcomes *across* a set of 20+ villages and also how within-village variation in social connectivity impacted risk *within* villages.

We employed multi-level data about common characteristics of individual villages, households, and the individuals within them. But at this time, I thought of myself as doing a few things, but none of them were spatial:

1. Infectious disease epidemiology: Why and how do people become *infected* with various pathogens?
2. Social epidemiology: How do social *relationships* impact disease outcomes?
3. Network analysis: How does the *structure* of relationships impact individual and community health?

As it happened, all of these things were correct. But what I didn't really understand at the time was that the collection of these different approaches into a single analysis made it spatial or geographic in nature, even if I didn't realize it

3.3 A process of progressive revelation

As a postdoc, working with Ted Cohen, I began analyzing data from a large study of household-level tuberculosis transmission in Lima, Peru. Figure 3.3 illustrates the model we developed to characterize household-level differences in transmission rates as a function of exposure type:

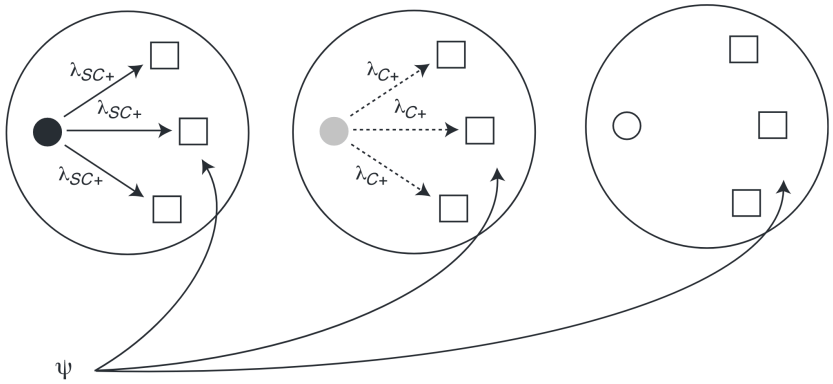


Figure 1. The figure illustrates the different sources of tuberculosis infection in the infection risk model. Smear-positive/culture-positive index cases (black circle) are hypothesized to be the most infectious, followed by smear-negative/culture-positive index cases (gray circle) and then smear-negative/culture-negative index cases (white circle). λ_{SC+} and λ_{C+} indicate the risk of infection for an uninfected household contact (white square) exposed to a smear-positive/culture-positive index case or smear-negative/culture-positive index case, respectively. ψ is the community risk of infection to which all households are subject. The solid arrows indicate a higher hypothesized level of infectiousness than values represented by the dashed arrows. Random intercepts are included in the model at the household and health center level to account for correlated responses within these units.

Figure 3.3: Characterizing household-level variation in risks of infection from community vs. household exposures (Figure from (5))

At the time, I knew that these households were distributed across neighborhoods of Lima, but I didn't give it much thought. I was more interested in risks experienced by an average

household. And to be honest, I didn't know that spatial metadata were available on each of the households, since I wasn't involved with the data collection!

In the interim, I got the chance to work on some collaborative projects with an explicitly spatial focus. In one, we reconstructed an outbreak of morbillivirus (think: measles) among a herd of migratory dolphins (Figure 3.4).

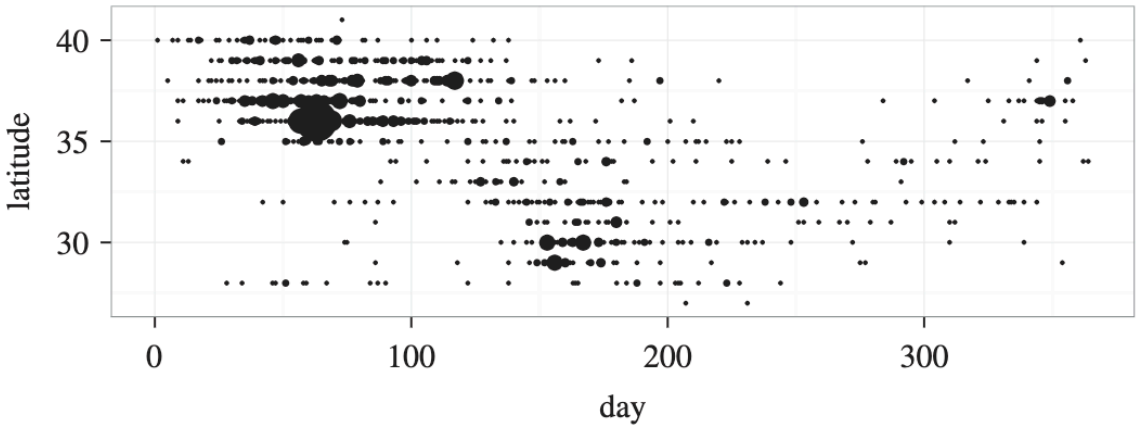


Figure 3.4: Locations of dolphin strandings during a morbillivirus outbreak in the North Atlantic (dot size indicates a greater number of strandings; Figure from (6))

In another, we looked at the relationship between environmental risks, such as neighborhood-level flooding, on the rate of pediatric diarrheal disease in Ho Chi Minh City in Vietnam (Figure 3.5).

These were the first experiences I had explicitly looking at these outcomes as a function of geographic space. While I had previously thought that mapping and spatial analysis and health geography were big scary things I couldn't do, I started to realize something important: These projects were not substantively very distinct from ones I had done before. The difference was that we were explicitly talking about spatial relationships and making maps (or simple one-dimensional diagrams as in Figure 3.4), instead of implicitly as in Figure 3.2 or Figure 3.3.

After completing these other projects, I dove back in to the Lima TB data to look at the drivers of multi-drug resistant TB (MDR-TB) risk. This was when I finally found out (some 2 years after I had started working with these data!) that spatial information on each household was available. So, with great trepidation, for the first time I made a map to explore spatial variability in MDR-TB outcomes.

And when I did this, we instantly saw that there were seemingly meaningful differences in the rate of TB overall, and MDR-TB in particular, across different health center catchment areas:

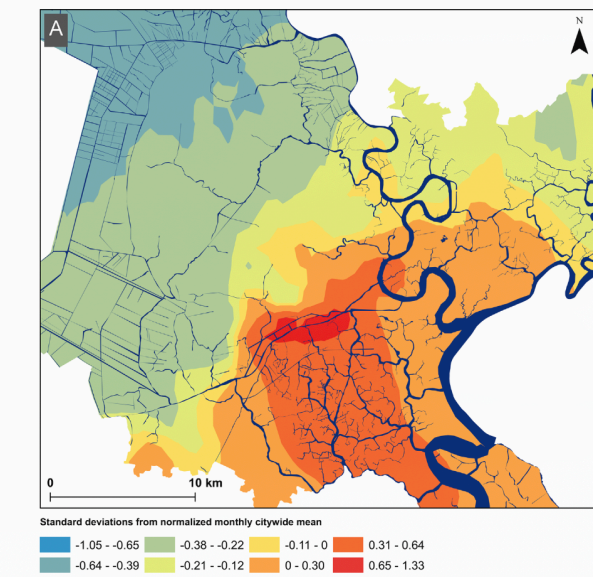


Figure 3.5: Incidence of pediatric diarrhea across neighborhoods of Ho Chi Minh City, Vietnam (Figure from (7))

This was the moment, some 10 years after I dipped my toes into the world of infectious disease epidemiology, where I realized I had been doing spatial work all along.

3.4 So what?

Why am I bothering you with this tedious and indulgent bit of personal history? *It's because it took me way too long to recognize that spatial epidemiology was a wrapper around a set of skills and ideas I had been working with for many years before I recognized what I was doing.* I was intimidated by anything preceded by 'spatial-': it sounded like a bunch of skills I didn't have and couldn't acquire.

My belated realization about the emergent quality of spatial epidemiology, and its broader connection to relational thinking in public health and the social sciences, has been crucially important for me. It made me realize that when I push into new areas - in life as much as research - that I probably have more of the tools I need than I realized in advance.

This means that you don't need to identify as a spatial or network or time series analyst to be one. And if you want to think of yourself as one, you should, because ultimately it is the intention to engage with the spatial, social and temporal relationships that drive health

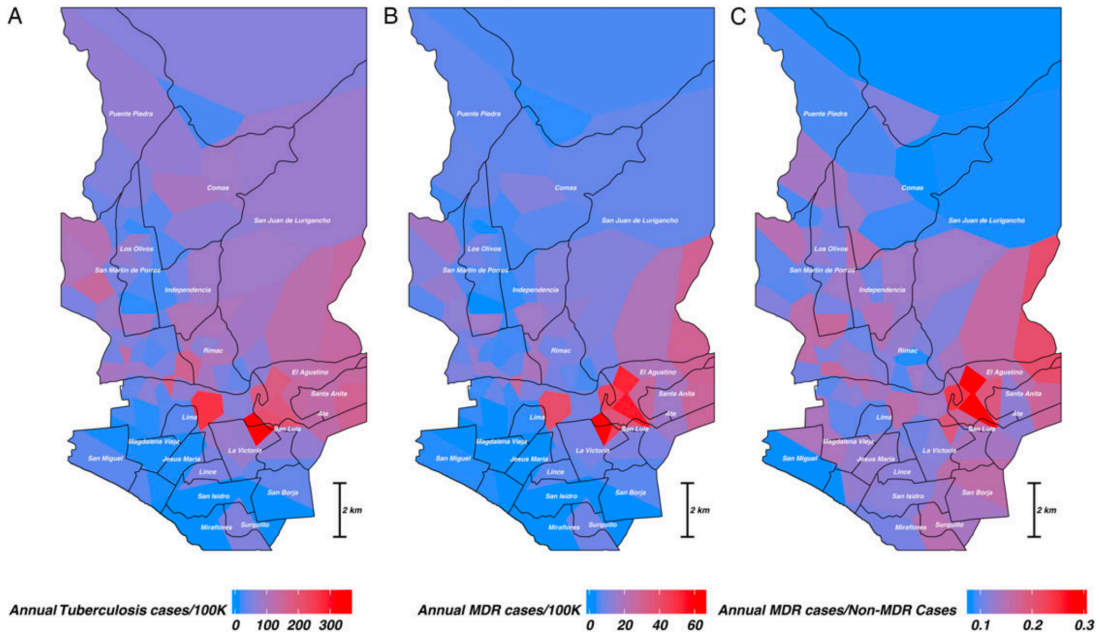


Figure 1. HC-level risks. Annual per-100 k rates of drug-sensitive and drug-resistant tuberculosis (*A*) and MDR tuberculosis (*B*), by HC catchment area. *C*, Ratio of the per-capita rate of MDR to non-MDR cases by HC. HC catchment areas are represented by polygons, with polygon fill color indicating the tuberculosis or MDR-tuberculosis rate in cases/100 K population. The boundaries of administrative districts of Lima are overlaid in black, and labeled in white**.** Abbreviations: HC, health center; MDR, multidrug-resistant.

Figure 3.6: The first map(s) I ever made, from (8), nearly 10 years after I started my research career.

outcomes that makes the difference. This is likely true for many if not most scientific subfields[1], but this one is mine and I'm glad I finally realized it!

References

[1]To be fair, you probably need to be working near-ish to a field for it to happen by chance: there is little chance of me taking on the characteristics a particle physicist or chemical engineer by chance, but I wouldn't rule out lepidopterist or archaeologist entirely!

4 What are the elements of relational epidemiology?

I arrived at the idea of relational epidemiology as an umbrella category from spatial and hierarchical analysis. The stereotypical challenge in this setting is to adjust away the impact of context to get at some more generally meaningful parameter estimate, e.g. a treatment effect. But another way of thinking about this is that the hard work is in characterizing not only the 'main' or fixed effects, but in capturing the drivers of variability in outcomes across locations.

But not every contextual question is strictly spatial: If we care about how the structure of social networks impacts individual and collective risks, we are talking about context once again. In the network example, the context is one's network 'neighborhood', the collection of individuals one interacts with. In reasonably homogeneous networks, groups of individuals may share very similar or almost identitical network contexts.

Ultimately, the key to understanding context is *relatedness*: Individuals sharing a context are likely to have similar relationships to their physical environment, the societies they are a part of, and potentially to each other, than those not sharing the same context. These relationships may be micro-level social relationships or macro-scale spatial ones, but they may also be temporal in nature. Temporal relationships may occur at a micro scale, e.g. the rise or fall in incidence of an infectious disease during a given week is likely to be a function of the prevalence of that same disease in the previous week. But these temporal relationships are also often more macro-scale and historical in nature: The long history of racial residential discrimination in the United States is undoubtedly a driver of many racial health dispartities we see today.

4.1 Problem-orientation is nonnegotiable

> "A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools." - Douglas Adams, Hitchhikers Guide to the Galaxy.

Sometimes, methods-y topics in epidemiology and public health are boiled down to a sequence of steps to be applied to each new problem. In the worst cases, they come to us as a series of

copy-paste, plug-and-chug pieces of code to be reused each time the same type of problem is encountered.

Beyond being a boring way to learn, this has the effect of putting the methods at the front of the train, with the question or problem implicitly assumed to be an opportunity or excuse to employ the model.

This is an approach that can get you some publications and maybe a little bit of clout within the musty world of academia, but it doesn't do much to solve the types of problems working epidemiologists face. And in truth, it doesn't even work so well on the academic side of the fence.

Inside the universe of this book, the problem is the first and most important thing. The question is the fixed point against which our analytic approaches are chosen. *This means that there are no fixed methodological answers to applied questions: Our methodological approaches and tools must be as diverse and heterodox as the questions the world throws at us.*

Making sense of the types of patterns we see in the real world requires us to first identify:

1. A question we want or need to answer.
2. The most important types of relationships impacting our outcome of interest (time, space, individual-to-individual).
3. A methodological approach that will allow us to characterize the impact of those relationships on the outcome we care about.

A methodological caboose

It is important to note that the choice of method comes *last* here: A key motivation of this book is to sidestep the tendency to train ourselves into methodological hammers looking for data nails to whack away at.

In addition to this, there is no pre-supposition that the appropriate method is necessarily 'fancy' in the sense of being conceptually or mathematically complex, computationally intensive, or even primarily or at all quantitative in nature. Whether this is the case is entirely a function of what happens at the intersection between question, data, and theory.

5 Some laws of relational and contextual epidemiology

Part II

Relationships

Part III

Space, time, and network

6 Social

6.1 Social Stratification & Inequality

6.2 Social Networks

7 Spatial

Please read the following two pieces that discuss key ideas about geospatial relatedness:

Miller HJ. Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*. 2004;94(2):284-289. doi:10.1111/j.1467-8306.2004.09402005.x

Goodchild MF. The Validity and Usefulness of Laws in Geographic Information Science and Geography. *Annals of the Association of American Geographers*. 2004;94(2):300-303. doi:10.1111/j.1467-8306.2004.09402008.x

7.1 Physical Space

7.2 Social Space

8 Temporal

8.1 Short-term correlation and fluctuation

8.2 Time series

8.3 History

Part IV

Methods

9 Codifying Tobler's First Law using Locally Weighted Regression

In this brief tutorial, we will review some basic ideas about smoothing and start thinking through how we can express these ideas mathematically and in R.

Tobler's profound - but deceptively simple - first law states that:

> "Everything is related to everything else. But near things are more related than distant things." (9)

He applied this idea to his development of a dynamic model of urban growth in the Detroit region which assumed that rates of population growth were spatially similar:

In this post, we're going to start with a simpler problem - change in the value of a function in one dimension - to see how we can translate the concept of *distance decay* implied by Tobler's first law (TFL) into a useful model. To begin, we're going to keep it simple with no noise or observation error and just interpolate some values.

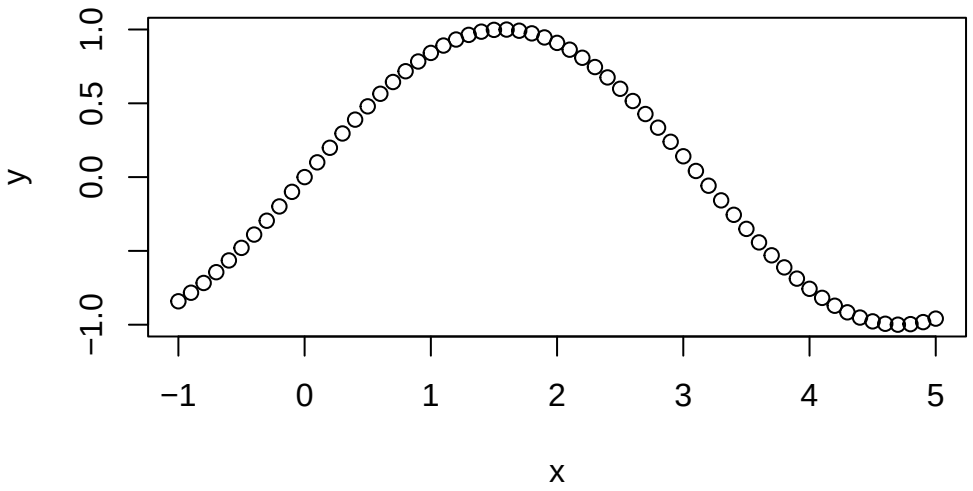
9.1 Notation and Terminology

In this example, we are interested in visualizing and predicting the values of a function $f(x_i)$ which outputs values y_i, the expected value of the output function.

$$y_i = f(x_i)$$

Lets start by getting the values of $f(x)$ for every input value x. For simplicity, we will assume that $f(x)$ is a sine function and that the values of x go from -1 to +5, allowing us to observe one half cycle of the sine function:

```
x <- seq(-1, 5, by = 0.1)
y <- sin(x)
plot(x,y)
```



You can see right away that this simple curve pretty neatly expresses Tobler's first law: $f(x)$ values of each point are in general more similar to each other for nearby values of x. If we want to press this idea into real-world practice, we need a *model* that can translate TFL into quantitative estimates and qualitative visualizations. There are lots of ways to do this, but we'll focus in on locally-weighted regression, also known LOWESS.

The basic idea of a LOWESS regression is to define a window of size k points around each value one wishes to estimate, and calculate a weighted average of the value of those points, which can then be used as the estimated value $\hat{y_j} \approx f(x_j)$. We then run the values of these nearest neighbors through a weight function $w(x)$.

These weight functions can take a lot of different forms, but we'll start simple with a uniform one, i.e. just taking the average of the k nearest neighbors, so that $\hat{y} = sum(z(x_i, k))/k$, where $KNNz$ is a function returning the y values of the k nearest observations to x_i. The value of k is sometimes referred to as the *bandwidth* of the smoothing function: Larger bandwidths use more data to estimate values at each point, smaller ones use less.

Making some locally weighted estimates

Using the `fnn` package for R, we can find the indices of the k nearest neighbors of each point we want to make an estimate at:

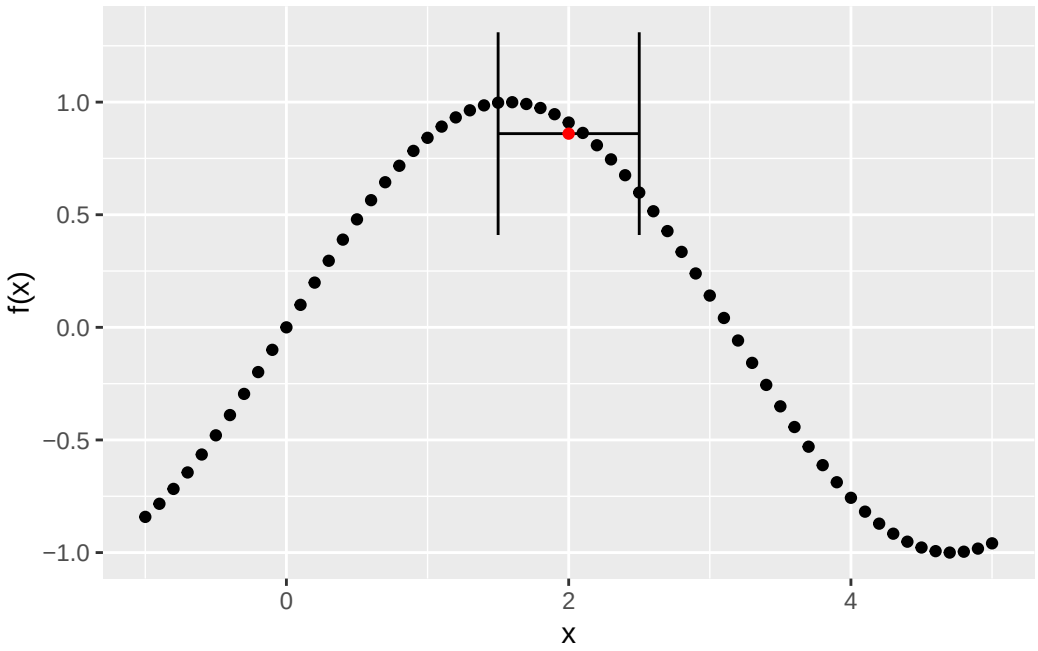
```
library(FNN)
k <- 10
z <- knn.index(x, k=k)
```

You can read the output of this function, below, as indicating the indices (i) of the 10 nearest points to each of the values of x.

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    2    3    4    5    6    7    8    9   10    11
 [2,]    1    3    4    5    6    7    8    9   10    11
 [3,]    2    4    1    5    6    7    8    9   10    11
 [4,]    5    3    6    2    1    7    8    9   10    11
 [5,]    6    4    7    3    8    2    1    9   10    11
 [6,]    5    7    4    8    3    9    2   10    1    11
 [7,]    8    6    9    5   10    4   11    3   12     2
 [8,]    7    9   10    6   11    5    4   12    3    13
 [9,]   10    8   11    7   12    6    5   13   14     4
[10,]    9   11    8   12    7   13   14    6    5    15
```

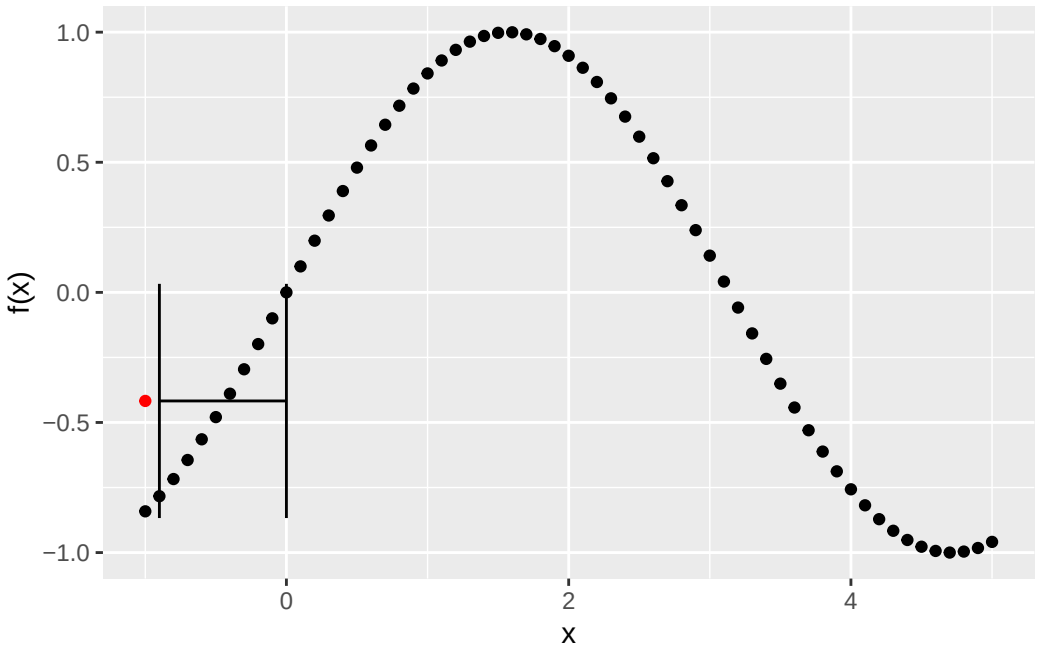
We can visualize this by picking a point in the middle of the series and its 10 nearest neighbors and the estimated value of $\hat{y}_i$ obtained by just taking the average of the k nearest points:

```
library(ggplot2)

## Plot the original data
g <- ggplot() + geom_point(aes(x=x, y = y)) +
  xlab("x") + ylab("f(x)")

## Now, get the index for x = 2
x_index <- which(x==2)

## Show the range of k nearest neighbors of this point
knn_low <- min(x[z[x_index, ]])
knn_high <- max(x[z[x_index, ]])
y_hat <- mean(y[z[x_index,]], )

## Add errorbars to the figure to show the 10 nearest values with the height of the point
g <- g + geom_errorbarh(aes(xmin = knn_low, xmax = knn_high, y = y_hat)) + geom_point(aes(

plot(g)
```



Notice that if the `knn` function is applied at the low end of the series, i.e. to the first value, it will use points to the right of that one instead of to either side:

```
library(ggplot2)

## Plot the original data
g <- ggplot() + geom_point(aes(x=x, y = y)) +
  xlab("x") + ylab("f(x)")

## Use the index for the lowest value
x_index <- 1

## Show the range of k nearest neighbors of this point
knn_low <- min(x[z[x_index, ]])
knn_high <- max(x[z[x_index, ]])
y_hat <- mean(y[z[x_index,]], )

## Add errorbars to the figure to show the 10 nearest values with the height of the point
g <- g + geom_errorbarh(aes(xmin = knn_low, xmax = knn_high, y = y_hat)) + geom_point(aes(

plot(g)
```

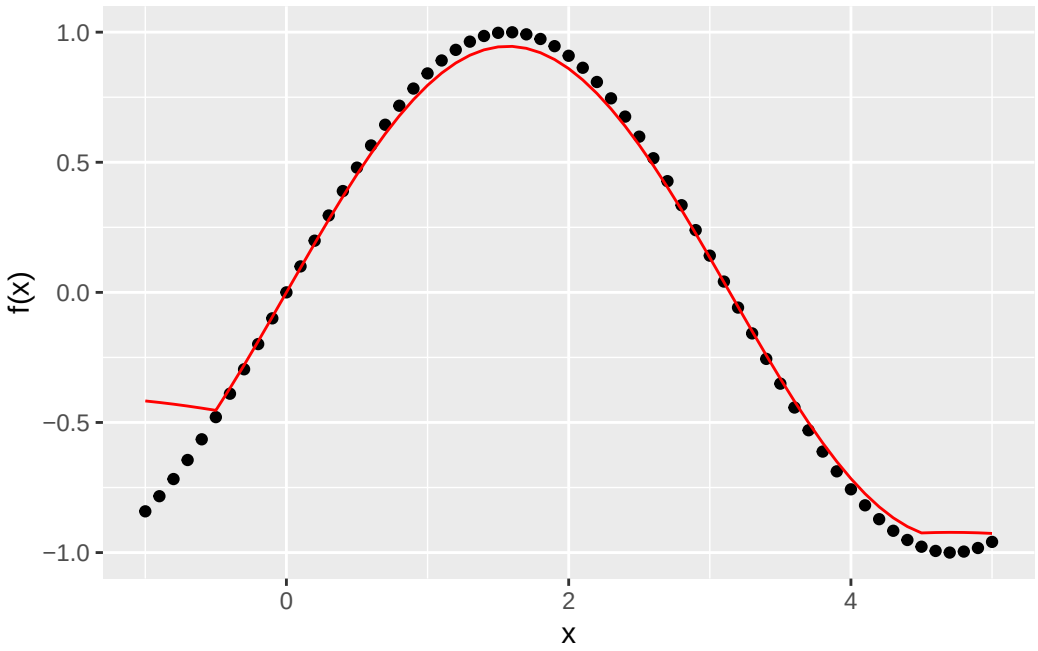


Now, lets see what happens if we run our smoother over the whole series and take the average of the 10 nearest points for each and compare them to the observed data:

```
y_hat <- rep(0, length(x))
for (i in 1:length(x) ) {
  y_hat[i] <- mean(y[z[i,]], )
}
```

Now plot the predicted vs. the observed values:

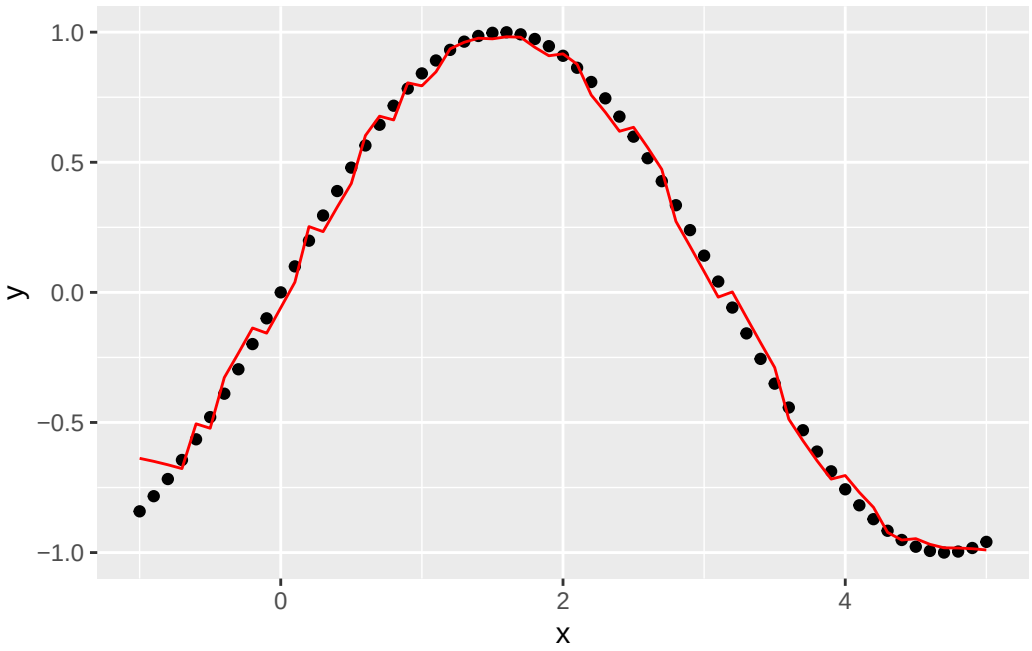
```
g <- ggplot() + geom_point(aes(x=x, y = y)) +
  xlab("x") + ylab("f(x)") + geom_line(aes(x=x, y=y_hat), colour = "red")
plot(g)
```



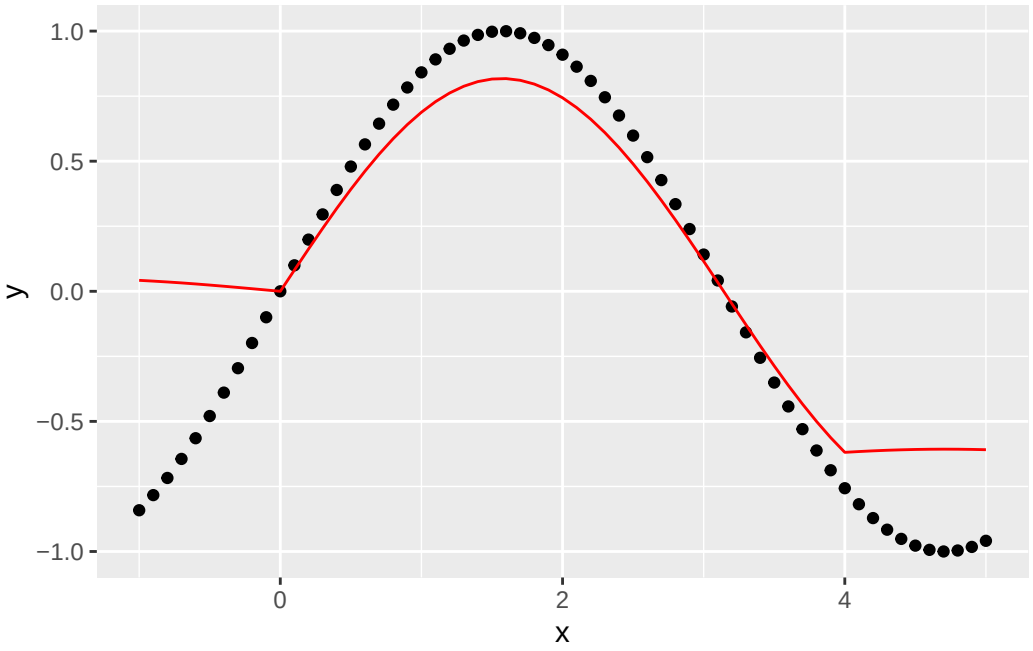
You can see this does a pretty good job all the way through, except at the edges. Lets try it again with a smaller window - or bandwidth - of 5 and see what happens. First, we'll write a function that will give us the predicted value of y at each point given a window of size k and an input value:

```
knn_est <- function(x, y, k) {
  z <- knn.index(x, k=k)
  y_hat <- rep(0, length(x))

for (i in 1:length(x) ) {
  y_hat[i] <- mean(y[z[i,]], )
}

  df <- data.frame(x=x, y = y, yhat = y_hat)
  return(df)
}
```

```
pred_df <- knn_est(x, y, 5)
g <- ggplot(pred_df) + geom_point(aes(x=x, y=y)) + geom_line(aes(x=x,y=yhat),colour="red")
plot(g)
```

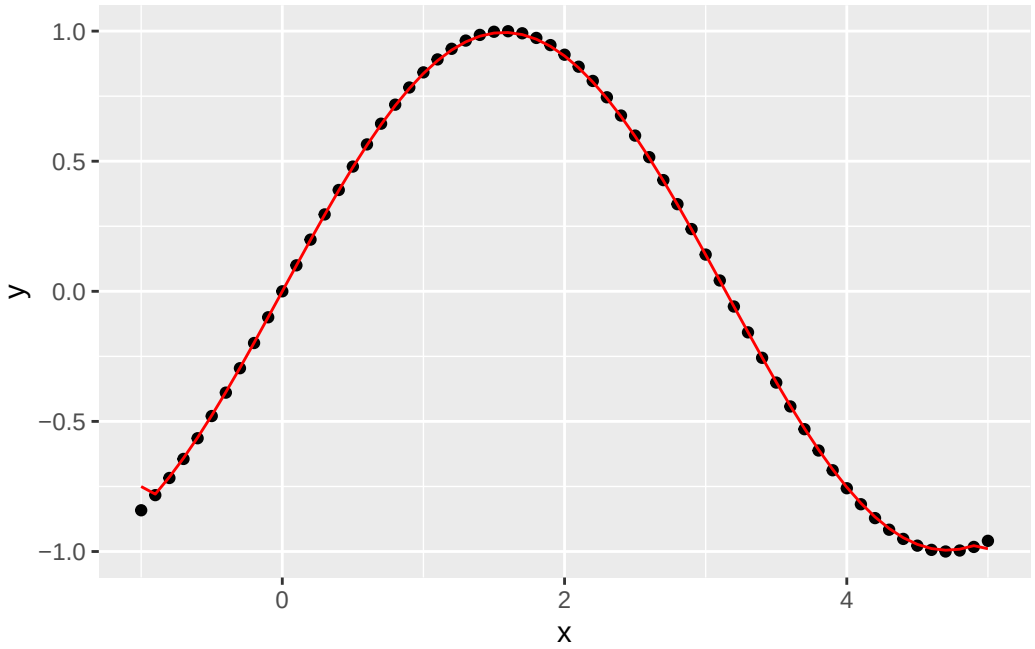


This gets rid of a lot of the weird effects at the edges but introduces some noise into the function. What if we make the window bigger, say 20, to get rid of some of the noise?



This seems to make the edge effects worse, as well as the estimates of the function overall worse.

What happens if we go in the opposite direction and shrink the window down to 2?



Discussion Questions

- Why does this appear to be more accurate for these data than $k = 10$ and $k = 5$?
- What would happen if we added observation noise to the values of y_i? Which one of the smoothers do you think would work better then?
- Is the one *best* value of k for all datasets? How might you go about picking the best one?
- How does our uniform weight function express Tobler's first law? What kind of weight function $w(x)$ might do a better job of capturing the notion of distance decay?

9.2 References

10 Spatial Density

When working with spatial data, our analytic task often falls into either of two rough categories.

1. *Estimating local averages of a continuous variable.* This could be something like neighborhood-by-neighborhood variation in average blood pressure or the intensity of some kind of environmental risk factor such as the intensity of fine (e.g. $PM_{2.5}$) dust which can cause respiratory illness.
2. *Estimating the local density of or incidence of a particular outcome.* For example, if we are trying to understand spatial variation in the incidence of a particular disease, we are interested in knowing how many cases of that disease are present in a given location or are likely to be present in some nearby unobserved location.

In this tutorial, we'll dig into the problem of spatial density estimation in one dimension along a spatial *transect*. The techniques discussed here form the basis of a number of approaches for cluster or hotspot analysis. For examples of smoothing local values of a continuous variable, see Chapter 9.

10.1 A motivating example

A spatial transect is an area of space along a line crossing a landscape. These are often used in ecology and forestry to assess the health of an environment, species diversity and other factors. Using a transect can help simplify the problem of spatial analysis down to one dimension rather than the usual two, while still providing a tremendous amount of useful information.

For example, (10) were interested in characterizing the intensity of exposure to triatomine bugs and other insect vectors of the pathogen *T. cruzi*, which causes Chagas disease.

Imagine we are estimating the density of some unknown insect vector along a 1 kilometer transect with the goal of characterizing the risk of infection with a vector-borne illness.

Figure 10.1: Example of an ecological transect from the US National Park Service (source)

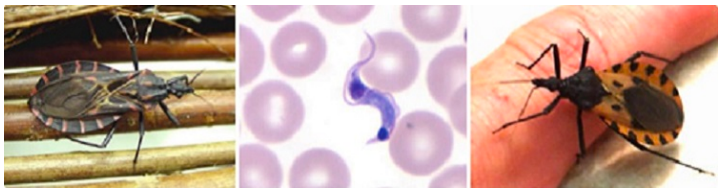


Figure 10.2: *Triatoma* (left- and right-hand panels) and *T. cruzi* (center) (source)

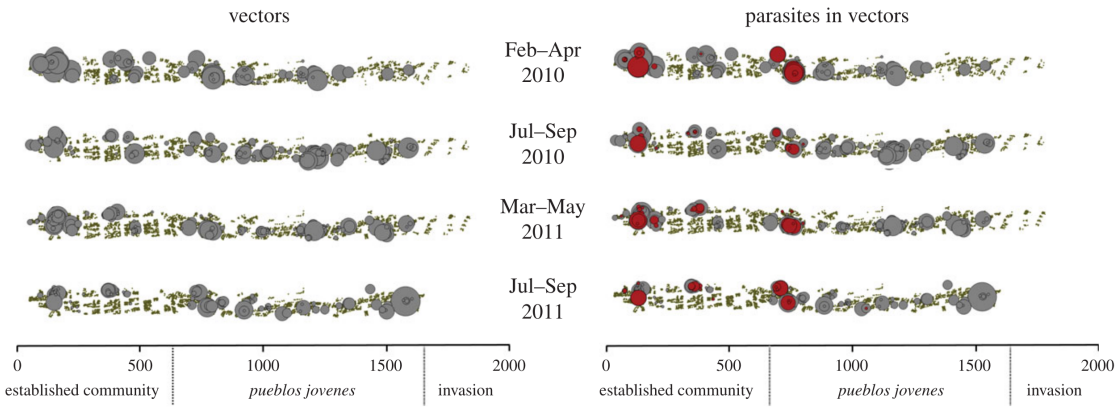


Figure 10.3: Intensity of *Triatomine* infestation along a 2km transect in Arequipa, Peru (Figure from (10))

10.2 Kernel density estimation in one dimension

Much like in our discussion of kernel smoothing of continuous outcomes, kernel functions play a key role in this setting as well. In this case, imagine that the locations of vectors along our transect have been sampled at random from some unknown function $f(x)$ which takes values from 0 (the beginning of the transect) to 1000m (the end).

We can use the Kernel function $K(d)$ to approximate the intensity of the outcome of interest at each observed case location x_i. Imagine that our observed data have locations $x_1, x_2, \ldots, x_n$ and that the distance between our point of interest, x_j and each observed point is $d_{ij} = |x_j - x_i|$.

Finally, lets include a bandwidth parameter, h, which controls the width of the window we will use for smoothing. When we put this all together, we can get an estimate of the density of our outcome of interest at location x_j as follows:

$$\hat{f}_h(x_j) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x_j - x_i}{h})$$

As you can see below, we can pick a range of kernel functions, but for the sake of simplicity, in this example, we will focus in on a Gaussian, or normal, kernel, which uses the probability density function of a normal distribution to weight points.

Lets start by sampling locations of observed points along a one dimensional line. To keep things interesting, we'll use a Gaussian mixture distribution with two components:

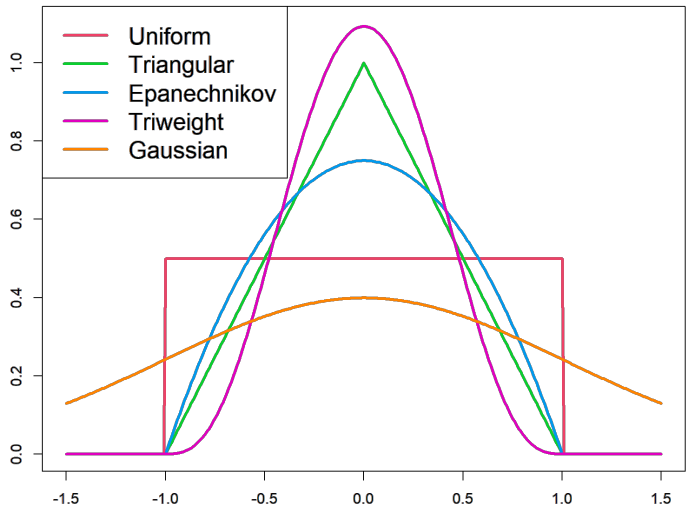


Figure 10.4: Comparison of different kernel functions (source)

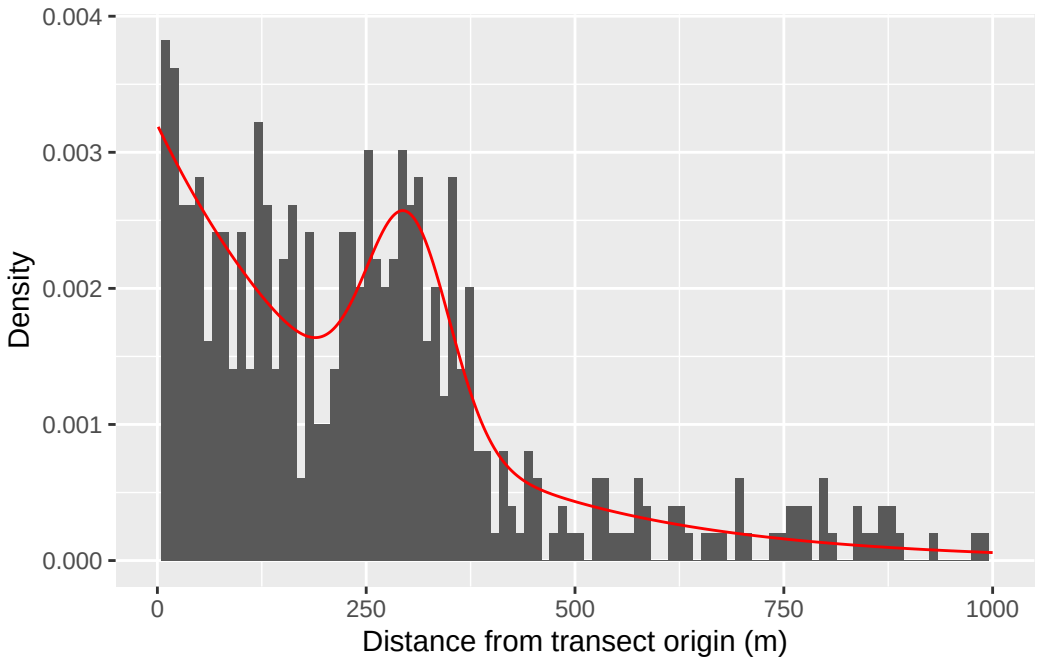
10.3 Worked example

First, lets imagine a scenario in which the risk of observing an insect vector steadily decreases as we walk along our transect. However, along the way there is a *hotspot* of increased risk beyond what we would expect from the smooth decline before and after that spot. For the purpose of this example, we'll assume that risk decays *exponentially* with distance from the origin, but that our hotspot is centered at a point 300 meters into the transect. The code below lets us sample the *locations* of the points along the transect where are observed from two distributions:

1. An exponential distribution representing smooth decay from the beginning to the end of the transect, and
2. A normal distribution representing a hotspot about 150m in width beginning 300m in

The figure below shows a histogram of locations sampled from $f(x)$ (vertical bars) overlaid with the true value of $f(x)$ in red:

```
library(ggplot2)
d_a <- dexp(1:1000, rate = 1/250)
d_b <- dnorm(1:1000, mean = 300, sd = 50)
y <- ((1-p_hot))*d_a + (p_hot*d_b)

dens_df <- data.frame(x = 1:1000, y = y)
xdf <- data.frame(x=x)


g <- ggplot(xdf) + geom_histogram(aes(x=x, y=..density..), bins=100) +
geom_line(data=dens_df, aes(x=x,y=y), colour="red") +
xlim(0, 1000) + ylab("Density") + xlab("Distance from transect origin (m)")
plot(g)
```



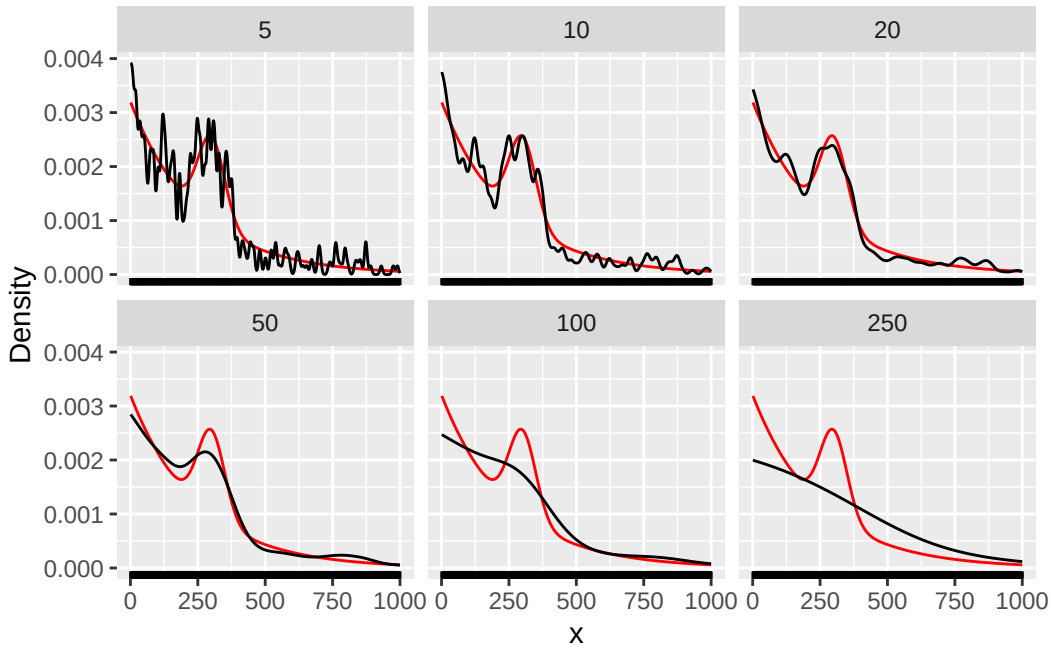
Now, imagine we have another set of finely spaced points along the line, and for each, we want to calculate the weight for each. The function below lets us do that:

The figure below shows the true value of our density function $f(x)$ in red, the density of points in the simulated data along the x-axis of the 'rug plot', and our smoothed density in black, for a bandwidth of $h = 10$:

```
library(ggplot2)
pred_df <- normal_smoother(x, h = 10)

g <- ggplot() + geom_rug(aes(x=x)) +
geom_line(data = pred_df, aes(x=x, y=y)) +
ylab("Density") + geom_line(data = dens_df, aes(x=x,y=y), colour="red") +
xlim(0, 1000)
dens_ojs <- dens_df
dens_ojs$y <- dens_ojs$y*cc
plot(g)
```

Now, lets see what happens if we try this for different values of h:

```
all_df <- data.frame()
for (hv in c(5, 10, 20, 50 ,100, 250)) {
  pred_df <- normal_smoother(x, h = hv)
  pred_df$h <- hv
  all_df <- rbind(all_df, pred_df)
}

  all_df$h <- as.factor(all_df$h)

g <- ggplot(all_df) + geom_rug(aes(x=x)) +
geom_line(data = dens_df, aes(x=x,y=y), colour="red") +
geom_line(aes(x=x, y=y)) +
ylab("Density") +
facet_wrap(~ h) +
xlim(0, 1000)

plot(g)
```




```
all_df <- data.frame()
hvals <- seq(1, 100, by = 2)
distvals <- seq(-100, 100, by = 1)
all_kernvals <- data.frame()
for (hv in hvals) {
  pred_df <- normal_smoother(x, h = hv)
  pred_df$h <- hv
  pred_df$smoother <- "gaussian"
  all_df <- rbind(all_df, pred_df)
  all_kernvals <- rbind(all_kernvals,data.frame(x=distvals, y=kdgaussian(distvals, bw = hv

  pred_df <- normal_smoother(x, h = hv, kern = kduniform, kernp=kpuniform)
  pred_df$h <- hv
  pred_df$smoother <- "uniform"
  all_df <- rbind(all_df, pred_df)
  all_kernvals <- rbind(all_kernvals,data.frame(x=distvals, y=kduniform(distvals, bw = hv)


  pred_df <- normal_smoother(x, h = hv, kern = kdtricube, kernp=kptricube)
  pred_df$h <- hv
  pred_df$smoother <- "tricube"
  all_df <- rbind(all_df, pred_df)
```

```
    all_kernvals <- rbind(all_kernvals, data.frame(x=distvals, y=kdtricube(distvals, bw =

  pred_df <- normal_smoother(x, h = hv, kern = kdtriangular, kernp=kptriangular)
  pred_df$h <- hv
  pred_df$smoother <- "triangular"
  all_df <- rbind(all_df, pred_df)
    all_kernvals <- rbind(all_kernvals, data.frame(x=distvals, y=kdtriangular(distvals, bw

}
all_df$y <- all_df$y * cc
```

10.4 Trying different bandwidths and kernels

You can adjust the range of the bandwidth here to get a better sense of the relationship between the smoothed curve (black) and true density (red). Adjust the bin width for the histogram of the underlying data to get a sense of the fit of the model to the underlying data.

```
// |echo: false
viewof h = Inputs.range([1, 100], {value: 10, step: 2, label: "Bandwidth (m)"})
viewof bw = Inputs.range([5, 100], {value: 10, step: 5, label: "Bin width (m)"})
viewof kern = Inputs.select(["gaussian", "uniform", "tricube", "triangular"], {value: "gau

numbins = Math.floor(1000/bw)

dtrans = transpose(hvals)
Plot.plot({
y: {grid: true,
label: "Density"},
x: {
label: "Distance from transect start (m) →"
},
    marks: [
    Plot.rectY(transpose(sample),Plot.binX({y: "count"}, {x: "loc", fill: "steelblue", thr
    Plot.lineY(dtrans, {filter: d => (d.h == h) && (d.smoother == kern), curve: "linear",
    Plot.lineY(transpose(dens), {x:"x", y: d => d.y * bw, curve:"linear", stroke: "red"})
    ]
})
```

The figure below shows the relative amount of weight placed on different points as a function of their distance from the point of interest (0, marked by the vertical red line):

```
// |echo: false
kv = transpose(kernvals).filter(d => d.h == h && d.smoother == kern)
Plot.plot({
y: {grid: true, label: "Relative weight of point as compared to origin"},
x: {
label: "Distance from point of interest (m)   "
},
marks: [
//Plot.lineY(kv, {filter: d => (d.smoother == kern), x:"x", y: d => d.y*1000}),
Plot.lineY(kv, Plot.normalizeY({x:"x", y: "y", basis: "extent"})),
Plot.ruleX([0], {stroke: "red"})
]})
```

Questions

- Which of the bandwidth options seems to do the best job in capturing the value of $f(x)$? Why?
- How does the choice of kernel impact the smoothing?
- How do the different kernel functions encode different assumptions about *distance decay*?
- What is the relationship between the histogram of the data and the smoother? What do you see as you change the histogram bin width relative to the smoothing bandwidth?

10.5 Additional Resources

Please see Matthew Conlen's excellent interactive KDE tutorial

References

1. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and STAN. 2nd edition. Boca Raton: Chapman and Hall/CRC; 2020.
2. Hilborn R, Mangel M. The Ecological Detective: Confronting Models with Data. 1st edition. Princeton, NJ: Princeton University Press; 1997.
3. Gelman A. Regression and Other Stories. 1st edition. Cambridge: Cambridge University Press; 2020.

4. Zelner JL, Trostle J, Goldstick JE, et al. Social Connectedness and Disease Transmission: Social Organization, Cohesion, Village Context, and Infection Risk in Rural Ecuador. *American Journal of Public Health* [electronic article]. 2012;102(12):2233–2239. (http://ajph.aphapublications.org/doi/10.2105/AJPH.2012.300795). (Accessed December 15, 2019)

5. Zelner JL, Murray MB, Becerra MC, et al. Age-Specific Risks of Tuberculosis Infection From Household and Community Exposures and Opportunities for Interventions in a High-Burden Setting. *American Journal of Epidemiology* [electronic article]. 2014;180(8):853–861. (https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwu192). (Accessed December 15, 2019)

6. Morris SE, Zelner JL, Fauquier DA, et al. Partially observed epidemics in wildlife hosts: Modelling an outbreak of dolphin morbillivirus in the northwestern Atlantic, June 2013–2014. *Journal of The Royal Society Interface* [electronic article]. 2015;12(112):20150676. (https://royalsocietypublishing.org/doi/10.1098/rsif.2015.0676). (Accessed December 15, 2019)

7. Thompson CN, Zelner JL, Nhu TDH, et al. The impact of environmental and climatic variation on the spatiotemporal trends of hospitalized pediatric diarrhea in Ho Chi Minh City, Vietnam. *Health & Place* [electronic article]. 2015;35:147–154. (https://linkinghub.elsevier.com/retrieve/pii/S1353829215001094). (Accessed December 15, 2019)

8. Zelner JL, Murray MB, Becerra MC, et al. Identifying Hotspots of Multidrug-Resistant Tuberculosis Transmission Using Spatial and Molecular Genetic Data. *Journal of Infectious Diseases* [electronic article]. 2016;213(2):287–294. (https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jiv387). (Accessed December 15, 2019)

9. Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* [electronic article]. 1970;46:234–240. (http://www.jstor.org/stable/143141). (Accessed January 13, 2022)

10. Levy MZ, Barbu CM, Castillo-Neyra R, et al. Urbanization, land tenure security and vector-borne Chagas disease. *Proceedings of the Royal Society B: Biological Sciences* [electronic article]. 2014;281(1789):20141003. (https://royalsocietypublishing.org/doi/full/10.1098/rspb.2014.1003). (Accessed January 23, 2023)