CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 12

Lecture 12A: Describing Temporal Trends in Data

Preface. Lecture 12A elaborates on some topics from Chapter 4 of "Statistics in Public Health" (1998) edited by Donna Stroup and Steven Teutsch. However, Lecture 12A does not constitute a summary of that Chapter.

Motivating example. Figure 4.2 on page 67 displays the daily numbers of children reporting to the ER in four Atlanta hospitals during the summer of 1993. Although Figure 4.2 reveals considerable day-to-day variation, there is some suggestion of a temporal trend: more ER visits took place at the beginning and end of the summer. Even so, the temporal trend is difficult to visualize from the raw numbers in Figure 4.2. This motivates us to estimate the temporal trend and represent it graphically.

Moving average approach. A simple approach to estimating the temporal trend is the calculation of a moving average. The simplicity of the moving average approach may not be evident from the general formula on page 68, but we can understand what is going on by writing out a few equations.

Let Y_t denote the number of ER visits on day t, and let M_t denote the underlying temporal trend at day t in the sense that $Y_t = M_t + E_t$ with E_t having expected value 0 and accounting for the day-to-day variation. The moving average approach estimates M_{10}, M_{20}, and M_{30} by

$$\widehat{M}_{10} := (Y_7 + Y_8 + Y_9 + Y_{10} + Y_{11} + Y_{12} + Y_{13})/7, \quad (1)$$

$$\widehat{M}_{20} := (Y_{17} + Y_{18} + Y_{19} + Y_{20} + Y_{21} + Y_{22} + Y_{23})/7, \quad (2)$$

and

$$\widehat{M}_{30} := (Y_{27} + Y_{28} + Y_{29} + Y_{30} + Y_{31} + Y_{32} + Y_{33})/7. \tag{3}$$

Equations (1) through (3) clarify how the temporal trend can be estimated in general. For any t we define $\widehat{M}_t$ as the average daily number of ER visits in the week surrounding day t. For days 1 to 3 we take the average over the first week; for days 90 to 92 we take the average over the last week. Figure 4.3 displays the estimated temporal trend in a solid curve.

Of course, the decision to define $\widehat{M}_t$ as the average daily number of ER visits in the week surrounding day t seems arbitrary. Why not take an average over fewer days (say, 3) or over more days (say, 21)? In fact, Figure 4.3 also shows (in a dashed curve) the estimated temporal trend based on an average over 21 days.

There are formal statistical criteria by which we may choose the number of days over which an average is taken, which statisticians refer to as a "bandwidth". Instead of discussing these criteria in detail, I will describe qualitatively the bias/variance tradeoff implicit in bandwidth selection.

Bias/variance tradeoff. To understand the bias/variance tradeoff implicit in bandwidth selection, consider two extreme strategies.

1. STRATEGY 1. Take the bandwidth equal to 92.

2. STRATEGY 2. Take the bandwidth equal to 1.

With STRATEGY 1 we would calculate $\widehat{M}_{10}$ as the average daily number of ER visits over all 92 days in the summer. But that is also how we would calculate $\widehat{M}_{20}$, $\widehat{M}_{30}$, and $\widehat{M}_t$ for any day t. Thus, we would be estimating the temporal trend as a horizontal line with height determined by the average daily number of ER visits over all 92 days in the summer.

With STRATEGY 2 we would define $\widehat{M}_{10}$ as Y_{10}. We would define $\widehat{M}_{20}$ as Y_{20}, $\widehat{M}_{30}$ as Y_{30}, and $\widehat{M}_t$ as Y_t for any day t. Thus, we would be estimating the temporal trend as a very bouncy curve that literally connected the dots in Figure 4.2.

While silencing distracting day-to-day fluctuations, STRATEGY 1 obliterates long-term patterns in the data such as elevated numbers of ER visits during the early and late parts of the summer.

Thus, STRATEGY 1 controls variability but at a terrible price in bias. Our estimation of the temporal trend as a horizontal line grossly and systematically understates ER visits in the early and late parts of the summer. We see the forest but not the trees.

On the other hand, STRATEGY 2 does not obliterate any long-term patterns, but they remain obscured by distracting day-to-day fluctuations.

Thus, STRATEGY 2 controls bias but at a terrible price in variance. Our estimation of the temporal trend as a bouncy curve prevents us from recognizing all but the most obvious long-term patterns. We see the trees but not the forest.

For the moving average approach to be useful, we must employ a less extreme strategy. We must tolerate a little variance in order to not have too much bias, and we must tolerate a little bias in order to not have too much variance.

As Figure 4.3 suggests, a nice tradeoff is achieved with a bandwidth of 21 days. The moving average based on a bandwidth of 21 days reveals peaks during the early and late parts of the summer as well as troughs around days 20 and 60. The moving average based on a bandwidth of 7 days also reveals these features but contains some artifacts. For instance, the minor spike at day 15 is most plausibly explained as a manifestation of day-to-day fluctuations.

Kernel smoothing approach. To motivate the kernel smoothing approach, recall that the moving average approach entailed taking

$$\widehat{M}_{10} := (Y_7 + Y_8 + Y_9 + Y_{10} + Y_{11} + Y_{12} + Y_{13})/7. \tag{4}$$

We can rewrite equation (4) as

$$\widehat{M}_{10} = w_{7;10}Y_7 + w_{8;10}Y_8 + w_{9;10}Y_9 + w_{10;10}Y_{10} + w_{11;10}Y_{11} + w_{12;10}Y_{12} + w_{13;10}Y_{13}, \tag{5}$$

where the "weights" $w_{7;10}$ through $w_{13;10}$ have the common value $1/7$.

Once we have rewritten equation (4) as equation (5), a question arises almost immediately: why are we assigning equal weights to Y_7 through Y_{13} when we want to estimate M_{10}? Shouldn't we assign the largest weight to Y_{10}, the next largest weights to Y_9 and Y_{11}, and the smallest weights to Y_7 and Y_{13}? Indeed, the assignment of different weights to Y_7 through Y_{13} is central to the kernel smoothing approach.

To carry out kernel smoothing, first we specify a "kernel function". One possible kernel function is the "triangular function"

$$K(x) := \frac{\max\{0, 4 - |x|\}}{16}, \tag{6}$$

so named because its graph produces a triangular shape. Then for each t we define $\widehat{M}_t$ by

$$\widehat{M}_t := \sum_{i=1}^{92} w_{i;t} Y_i, \tag{7}$$

where the weights $w_{i;t}$ are determined by

$$w_{i;t} := K(i - t). \tag{8}$$

To illustrate, let us compute $\widehat{M}_{10}$. We have $w_{7;10} = K(-3) = 1/16$, $w_{8;10} = K(-2) = 2/16$, $w_{9;10} = K(-1) = 3/16$, $w_{10;10} = K(0) = 4/16$, $w_{11;10} = K(1) = 3/16$, $w_{12;10} = K(2) = 2/16$, and $w_{13;10} = K(3) = 1/16$.

Moreover, $w_{i;10} = 0$ for any i less than 7 or greater than 13 because $K(x) = 0$ for any integer x less than -3 or greater than 3. Thus, through kernel smoothing we obtain

$$\widehat{M}_{10} = \frac{1}{16}Y_7 + \frac{2}{16}Y_8 + \frac{3}{16}Y_9 + \frac{4}{16}Y_{10} + \frac{3}{16}Y_{11} + \frac{2}{16}Y_{12} + \frac{1}{16}Y_{13}. \quad (9)$$

We can similarly obtain

$$\widehat{M}_{20} = \frac{1}{16}Y_{17} + \frac{2}{16}Y_{18} + \frac{3}{16}Y_{19} + \frac{4}{16}Y_{20} + \frac{3}{16}Y_{21} + \frac{2}{16}Y_{22} + \frac{1}{16}Y_{23} \quad (10)$$

and

$$\widehat{M}_{30} = \frac{1}{16}Y_{27} + \frac{2}{16}Y_{28} + \frac{3}{16}Y_{29} + \frac{4}{16}Y_{30} + \frac{3}{16}Y_{31} + \frac{2}{16}Y_{32} + \frac{1}{16}Y_{33}. \quad (11)$$

Note that the kernel function has been specified in such a way that, for any t, the weights used to calculate $\widehat{M}_t$ sum to 1.

LOESS smoothing approach. The next step up from kernel smoothing is LOESS smoothing, which you have already encountered in Lecture 2 of CPH 931. LOESS smoothing would define $\widehat{M}_{10}$ in accordance with the solution to a weighted least squares problem.

Specifically, we would assert the existence of coefficients α_{10} and β_{10} such that, for t close to 10,

$$M_t \approx \alpha_{10} + \beta_{10} \times t. \quad (12)$$

Our estimates of α_{10} and β_{10}, call them $\hat{\alpha}_{10}$ and $\hat{\beta}_{10}$, would be the values of $r_{0;10}$ and $r_{1;10}$ that minimized the weighted sum of squares

$$\sum_{i=1}^{92} w_{i;10}[Y_i - r_{0;10} - r_{1;10} \times i]^2. \quad (13)$$

Note that expression (13) could be rewritten as

$$w_{7;10}[Y_7 - r_{0;10} - r_{1;10} \times 7]^2 + w_{8;10}[Y_8 - r_{0;10} - r_{1;10} \times 8]^2 + w_{9;10}[Y_9 - r_{0;10} - r_{1;10} \times 9]^2$$

$$+w_{10;10}[Y_{10} - r_{0;10} - r_{1;10} \times 10]^2 + w_{11;10}[Y_{11} - r_{0;10} - r_{1;10} \times 11]^2$$
$$+w_{12;10}[Y_{12} - r_{0;10} - r_{1;10} \times 12]^2 + w_{13;10}[Y_{13} - r_{0;10} - r_{1;10} \times 13]^2.$$

Then we would define $\widehat{M}_{10}$ as $\hat{\alpha}_{10} + \hat{\beta}_{10} \times 10$. We would obtain $\widehat{M}_{20}$, $\widehat{M}_{30}$, etc., in analogous fashion.

To understand why LOESS smoothing is a step up from kernel smoothing, consider replacing the "local linear model" from equation (12) with the "local constant model"

$$M_t \approx \alpha_{10} \tag{14}$$

for t close to 10. The corresponding weighted least squares problem is to minimize

$$\sum_{i=1}^{92} w_{i;10}[Y_i - r_{0;10}]^2. \tag{15}$$

Our estimate of α_{10}, call it $\hat{\alpha}_{10}$, would be the value of $r_{0;10}$ that minimized expression (15).

A standard calculus argument can be used to show that expression (15) is minimized when

$$r_{0;10} = w_{7;10}Y_7 + w_{8;10}Y_8 + w_{9;10}Y_9 + w_{10;10}Y_{10} + w_{11;10}Y_{11} + w_{12;10}Y_{12} + w_{13;10}Y_{13}.$$

Hence, we set

$$\hat{\alpha}_{10} := w_{7;10}Y_7 + w_{8;10}Y_8 + w_{9;10}Y_9 + w_{10;10}Y_{10} + w_{11;10}Y_{11} + w_{12;10}Y_{12} + w_{13;10}Y_{13}, \tag{16}$$

and $\widehat{M}_{10}$ is defined accordingly.

Thus, kernel smoothing is really a special case of LOESS smoothing when a local constant model is used instead of a local linear model. Moreover, the moving average is really a special case of LOESS smoothing when a local constant model is used and all nonzero weights are equal.

Lecture 12B: Discovering Spatial Patterns in Data

Preface. Lecture 12B borrows heavily from materials that Dr. Branscum prepared for his class on spatial statistics. At the risk of stating the obvious, Lecture 12B is not a substitute for a course on spatial statistics!

Motivation. Characterizing spatial distributions of disease events, and identifying factors predictive of where disease occurs, are core components to epidemiologic and public health practice and research. We want to identify and understand patterns in disease occurrence because understanding where events occur can provide information about why they occur, which is a step toward controlling disease.

Historical example of spatial data analysis. John Snow's study of the cholera epidemic in 1854 London is a good historical example of spatial data analysis. Snow hypothesized that cholera was being transmitted through drinking water. His hypothesis went against the popular theory at the time, which was that diseases such as cholera were caused by pollution. No one really knew the underlying mechanism by which the disease was transmitted, but Snow did not believe that it was due to breathing foul air.

Cholera deaths appeared to cluster around the Broad Street pump (Figure 1), but some such clustering might have been expected since many people lived in areas close to public drinking water supplies. In other words, some clustering could have been attributable to the greater population density rather than to a phenomenon involving the drinking water. Thus, the real question was whether the observed clustering exceeded what would have been expected given the population density.

Snow categorized cholera deaths according to one of two water companies that serviced the area. Careful control selection (i.e., identification of local people who did not have cholera) and extensive surveying of households with at least one cholera death, along with the use of maps, led to wider support of Snow's theory.

FIGURE 1

Challenges of spatial data analysis. Some statistical methods for spatial data analysis are related to, or are modifications of, familiar tools such as logistic regression and Poisson regression. However, spatial data analysis is challenging because the data are correlated rather than independent.

As simply put by Waldo Tobler in his first law of geography, "Everything is related to everything else, but near things are more related than far things." This statement reflects the notion that pairs of observations collected from nearby locations tend to be more alike than pairs collected from locations that are far apart.

More explicitly, suppose that $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ represent coordinates for latitude and longitude. Let Y_{u_i,v_i} denote the actual count of a disease event in the vicinity of latitude u_i and longitude v_i, and suppose that Y_{u_i,v_i} can be represented as $M_{u_i,v_i} + E_{u_i,v_i}$, where M_{u_i,v_i} is the expected count and E_{u_i,v_i} is a random fluctuation.

We anticipate that M_{u_i,v_i} will be close to M_{u_j,v_j} when (u_i, v_i) is close to (u_j, v_j). That by itself is not any problem. In fact, with Poisson regression we would model the logarithm of M_{u_i,v_i} as a polynomial in u_i and v_i. Since polynomials are continuous functions, Poisson regression would capture the phenomenon of M_{u_i,v_i} being close to M_{u_j,v_j} when (u_i, v_i) is close to (u_j, v_j).

The difficulty is that E_{u_i,v_i} will also be close to E_{u_j,v_j} when (u_i, v_i) is close to (u_j, v_j). In other words, the random fluctuations are correlated, which defies the assumption of independent data that underlies most of our familiar tools.

Spatial data analysis methods must take into account such correlations to yield valid estimates and meaningful conclusions.

Types of spatial data. Point referenced data, also called geocoded or geostatistical data, entail the collection of information at pre-specified locations. For instance, we may gather information on particulate matter at each of several pre-specified sites for monitoring air pollution.

Areal data, also called lattice data, involve the collection of information within well-defined geographic units. Such information is often recorded at the level of the geographic units themselves rather than for specific individuals living in the geographic units. Hence, areal data are often used for ecological studies. For instance, we may gather information on cancer incidence within a county, state, or census tract.

Point pattern data are, in a sense, the opposite of point referenced data. Instead of counting random numbers of disease events at pre-specified locations, we label the random locations at which disease events occur. An illustration would be a map of 1854 London showing a dot at every household in which someone died of cholera.

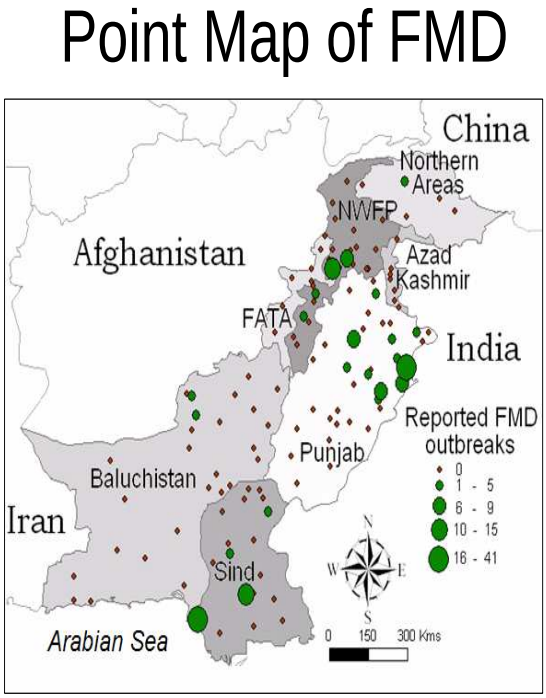
Sources of spatial data. Much effort has been devoted to gathering spatial data and making those data freely available online. Spatial data are often maintained by governmental agencies, such as local and state health departments.

More specifically, spatial data can come from vital statistics records, disease registries, health surveys, surveillance activities, census records, climate monitoring, and aerial photography.

Given the diversity of sources of spatial data, one anticipates that there may be some hurdles to overcome in data analysis besides the statistical issue of correlated errors. Such hurdles include the underreporting of disease events and the potential for mismatching types of spatial data. The former occurs when cases are undiagnosed, whereas an example of the latter

is having areal data on a health outcome but point referenced data on an environmental exposure thought to be related to the health outcome.

FIGURE 2



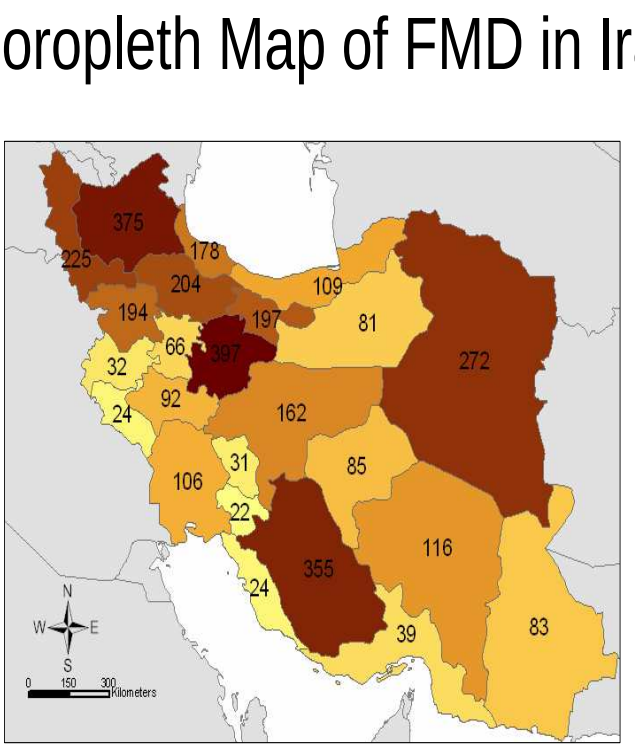
Visualizing spatial data. There are several ways to construct maps of spatial data. Two of the most common are point maps and choropleth maps.

Point maps are used with point referenced data. Figure 2 is illustrative: small red points indicate no reported outbreaks of foot-and-mouth disease

(FMD), small green points indicate between 1 and 5 reported outbreaks, and progressively larger green points indicate larger numbers of reported outbreaks.

Choropleth maps are used with areal data. Figure 3 is illustrative: regions in Iran with relatively few FMD cases have lighter shades of brown than regions with relatively many FMD cases.

FIGURE 3



Smoothing spatial data. In Lecture 12A we talked about smoothing temporal data. Often there is interest in smoothing spatial data as well. To understand why, consider the following (admittedly artificial) example.

Suppose that in the year 2008 there are 5 events in a region containing 5 million people. Suppose, moreover, that the region is divided into 5000 subregions each containing 1000 people. Then, even though the event rate for the region is 1/1000000, one or more subregions will have event rates of at least 1/1000 – that is 1000 times the event rate for the region! Moreover, at least 4995 subregions will have an event rate of 0.

Although the numbers for the year 2008 are what they are, they can be misleading in the following sense: we do not anticipate that the risk of a person getting the disease in the year 2009 is going to be as high as 1/1000 or as low as 0 in any subregion.

Spatial smoothing borrows information from neighboring subregions to produce better estimates of the risks for a future time period.

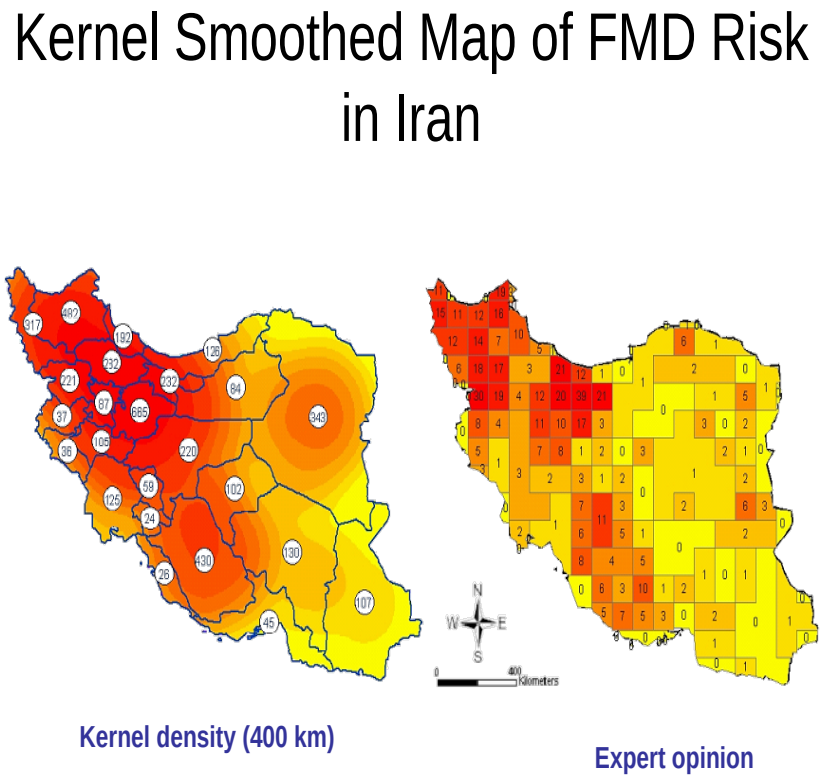
Locally weighted averaging. A common method for spatial smoothing, analogous to the moving average for temporal data in Lecture 12A, is the locally weighted average. Instead of displaying the observed event rate $r(u_i, v_i)$ for the region with latitude u_i and longitude v_i, we display

$$\sum_{j:(u_i-u_j)^2+(v_i-v_j)^2\leq\delta} n_i^{-1} r(u_j, v_j),$$

where δ is a positive constant that is analogous to a bandwidth and n_i is the number of indices j for which $(u_i - u_j)^2 + (v_i - v_j)^2 \leq \delta$.

Kernel smoothing. The locally weighted average for spatial data can be generalized to provide larger weights for some observations than others, just as was true of the moving average for temporal data. The n_i^{-1} is replaced by nonnegative weights $w_{i,j}$ such that $\sum_{j:(u_i-u_j)^2+(v_i-v_j)^2\le\delta} w_{i,j} = 1$ and $w_{i,j}$ is largest when $(u_i - u_j)^2 + (v_i - v_j)^2$ is smallest. Thus, observations from nearer regions receive larger weights. Figure 4 provides an illustration.

FIGURE 4



Discussion Questions

1. In Lecture 12A I stated that a kernel function should be such that the weights for $\widehat{M}_t$ sum to 1. To understand why this is important, suppose that we had instead defined

$$K(x) := \max\{0, 4 - |x|\}. \tag{17}$$

   This would yield, for instance,

$$\widehat{M}_{10} = Y_7 + 2Y_8 + 3Y_9 + 4Y_{10} + 3Y_{11} + 2Y_{12} + Y_{13}. \tag{18}$$

   What is the problem here?
   Hint: From Figure 4.2 we see that (rounding to the nearest ten) $Y_7 = 460$, $Y_8 = 460$, $Y_9 = 440$, $Y_{10} = 390$, $Y_{11} = 390$, $Y_{12} = 410$, and $Y_{13} = 480$.

2. In Lecture 12B I identified some challenges to spatial data analysis. Can you identify some similar (or different) challenges to temporal data analysis?