UNIVERSITY of WASHINGTON

Data Science UW Methods for Data Analysis

Time Series, Spatial Stats, and Intro to Bayesian Stats
Lecture 7
Nick McClure

The spatial world according to Twitter

Topics

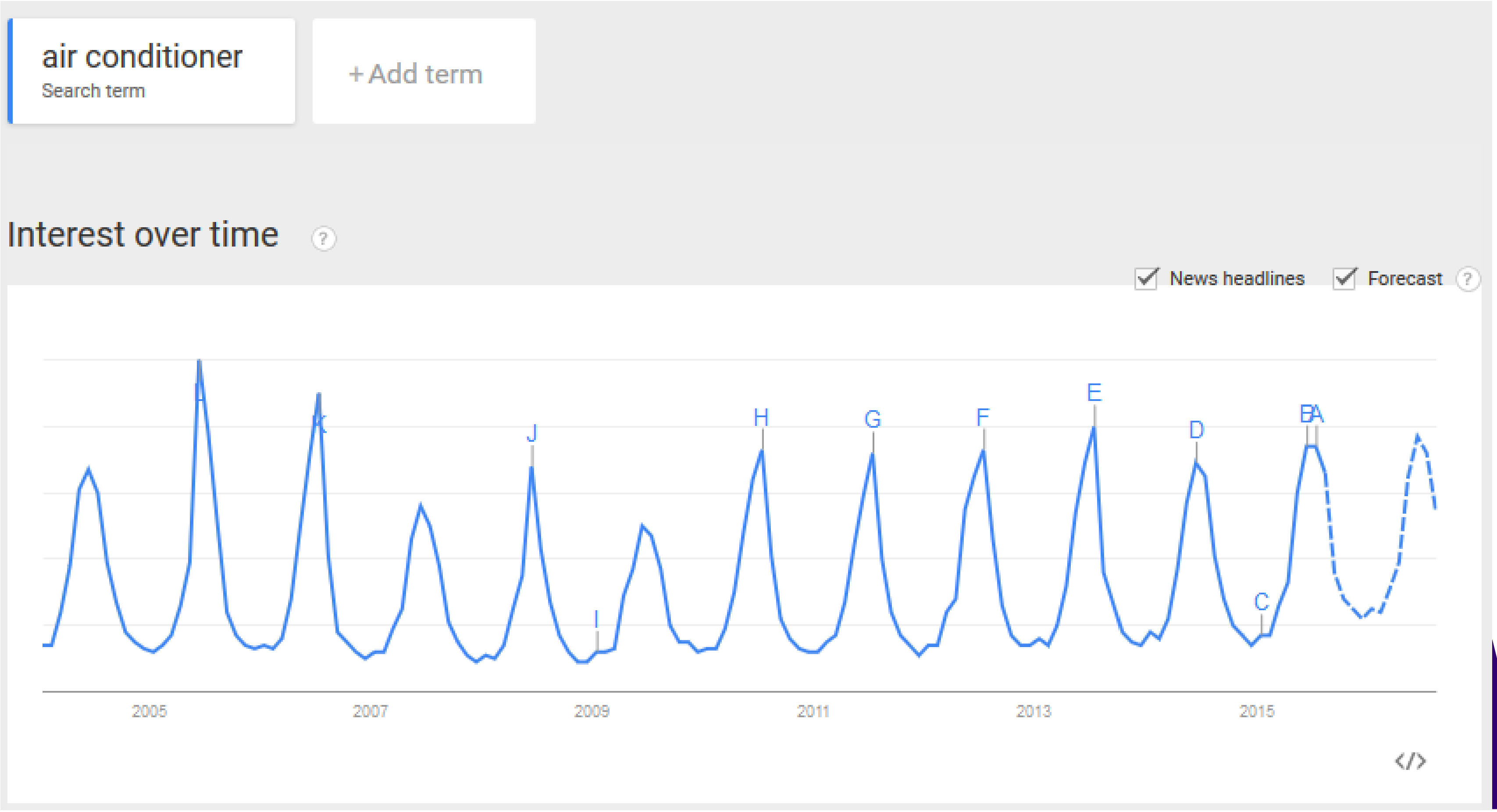
- Review
- Time series
- Spatial statistics
- Introduction to Bayesian Statistics

Review

- Decomposition methods
- SVD
  - SVD as linear regression
  - Variable reduction
  - Storing data
- Ridge Regression
- Lasso Regression
- Logistic Regression
- Binary classification
- Intro to time series

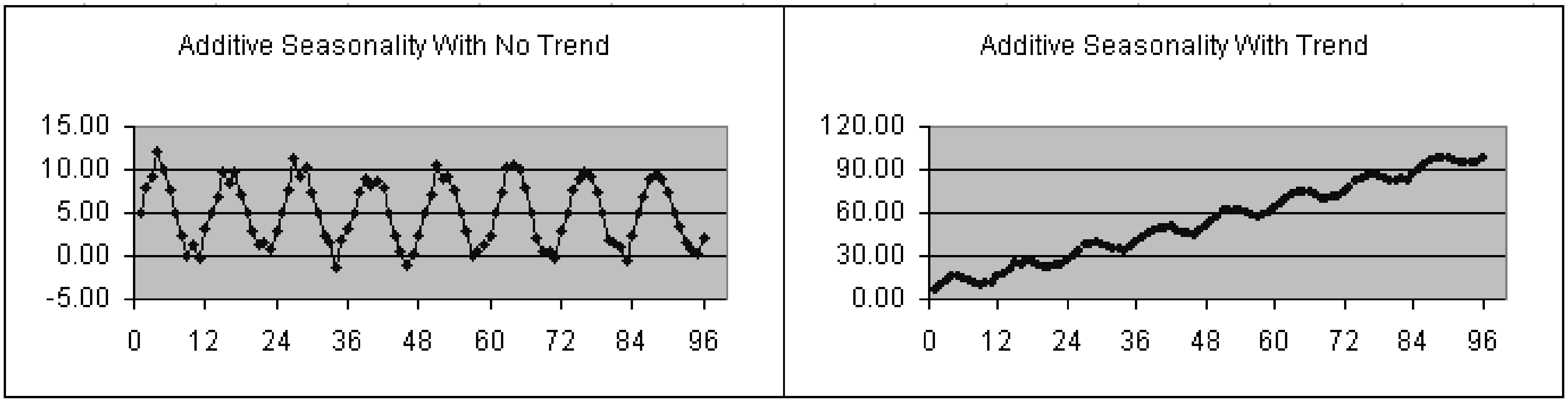
Time series

> How do we detect seasonality?



Time series

- Seasonality can be with or without trend
- If without trend, the series is called stationary



The Fourier Transform

- The Fourier transform maps a function (or series of points) to the frequencies that make up the function.
- It does this by averaging the normalized points across certain frequencies.

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{i2\pi k\frac{n}{N}}$$

- **To find the energy at a particular frequency, spin your signal around a circle at that frequency, and average a bunch of points along that path.**

Exponential Smoothing

> Past View moving average in which observations are weighted in terms of recency.

$$s_0 = x_0$$

$$s_t = \alpha x_t + (1-\alpha)s_{t-1}$$

$$s_0 = x_0$$

$$s_1 = \alpha x_1 + (1-\alpha)s_0 = \alpha x_1 + (1-\alpha)x_0$$

$$s_2 = \alpha x_2 + (1-\alpha)s_1 = \alpha(x_2 + (1-\alpha)x_1) + (1-\alpha)^2 x_0$$

$$s_2 = \alpha(x_3 + (1-\alpha)x_2 + (1-\alpha)^2 x_1) + (1-\alpha)^3 x_0$$

$$coefficients = \{1, (1-\alpha), (1-\alpha)^2, (1-\alpha)^3, \ldots\}$$

> This coefficient sequence is geometric progression, which is a discrete exponential function.

Double Exponential Smoothing

- Exponential smoothing does not do well with trends.
- To compensate for trends, we just add a term in describing the change between adjacent points.

$$s_0 = x_0$$

$$s_1 = \alpha x_1 + (1-\alpha)s_0$$

$$b_1 = x_1 - x_0$$

$$s_t = \alpha x_t + (1-\alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}$$

Triple Exponential Smoothing

> Triple exponential smoothing takes into account seasonality, or a tendency for the series to repeat itself.
> We are still accounting for linear trend as well.
> How do we find the length of the cycle?
  – Auto cross correlation methods, like the Fourier transform.

> R-demo

Autoregressive Model (AR)

> If a series is stationary and auto-correlated, it should be able to be predicted as some multiple of previous values.

> Every new observed point relies on what the previous p-points were:

$$y_t = c + \sum_{i=1}^{p} (\varphi_i y_{t-i}) + \varepsilon_t$$

> The above is shown as AR(p)

> R-demo

Auto Regressive Moving Average (ARMA)

- Auto-Regressive Moving Average (ARMA)
- ARMA is denoted by two variables (P,Q)
  - P = Auto regression order
  - Q = Order of moving average

$$y_t = c + \sum_{i=1}^{P} (\varphi_i y_{t-i}) + \sum_{i=1}^{Q} (\theta_i \varepsilon_{t-i}) + \varepsilon_t$$

AR (P,Q)= AR filter + MA filter + error terms

- R-demo

ARIMA

- Auto-Regressive Integrated Moving Average (ARIMA)
- ARIMA models are designated by three parameters:
- P = Order of Auto regression
- D = Degree of Difference for the ‘integrated’ part, this is how the model takes into account the differences needed for finding trend.
- Q = Order of Moving Average

- *Note ARIMA(0,0,0)=> $y_t = \varepsilon_t$ (random noise)

ARIMA(P,D,Q)=AR filter + Integration Filter + MA filter + error terms

(Long term)+ (stochastic trend) +(short term)+ error

AR Models

> ARIMA(1,0,0) = 1st order auto regressive

$$y_t = c + \varphi_1 y_{t-1} + \varepsilon_t$$

> ARIMA(2,0,0) = 2nd order auto regressive

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t$$

> Note that these models look very similar to random walks. This is because as the coefficients approach 1 they are the same.

> Consider these models similar to random walks, but not 'as-dependent' on the previous values.

ARIMA(P,D,Q)=AR filter + Integration Filter + MA filter + error terms

(Long term)+ (stochastic trend) +(short term)+ error

MA Models

- ARIMA(0,0,1) = 1st order Moving Average

$$y_t = c + \varphi_1 y_{t-1} + \varepsilon_t$$

- ARIMA(0,0,2) = 2nd order Moving Average

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t$$

- Note that these models look very similar to random walks. This is because as the coefficients approach 1 they are the same.
- Consider these models similar to random walks, but not 'as-dependent' on the previous values.

ARIMA(P,D,Q)=AR filter + Integration Filter + MA filter + error terms

(Long term)+ (stochastic trend) +(short term)+ error

Integrated Models (Random Walks)

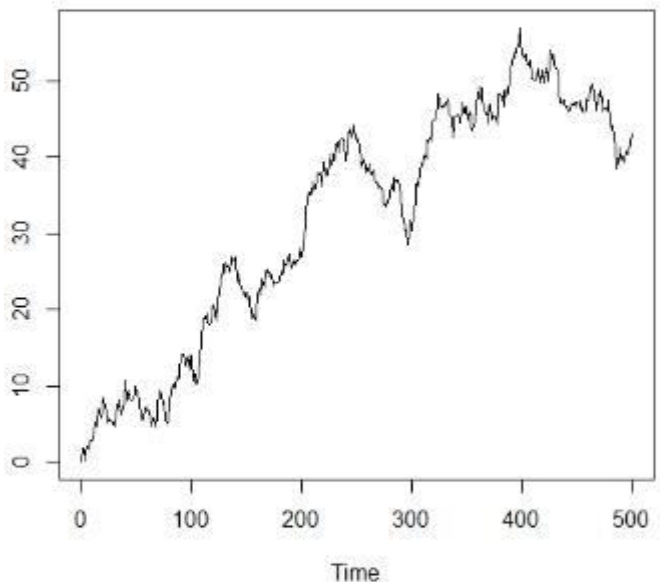
> ARIMA(0,1,0) = Random Walk Model

$$y_t = y_{t-1} + \varepsilon_t \quad \text{OR} \quad \Delta y_t = \varepsilon_t$$

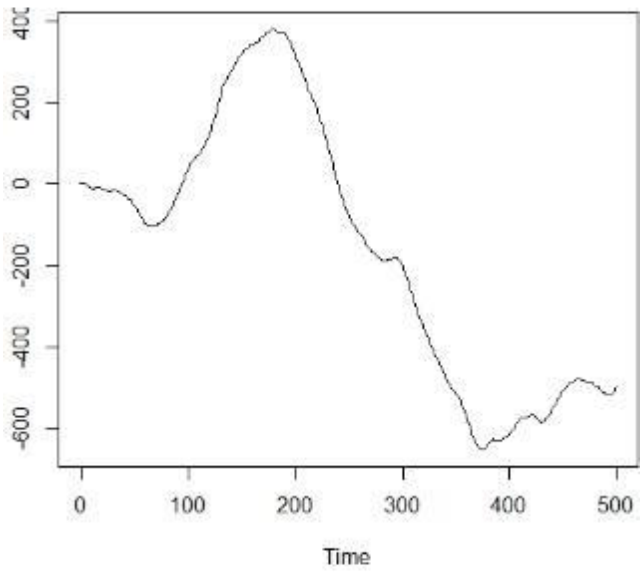
> ARIMA(0,2,0) = 2nd order random walk

$$y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) + \varepsilon_t$$

1st order



2nd order



ARIMA

ARIMA(P,D,Q)=AR filter + Integration Filter + MA filter + error terms

(Long term)+ (stochastic trend) +(short term)+ error

> ARIMA(1,0,0) = 1st order autoregressive

> ARIMA(0,0,1) = 1st order moving average

> ARIMA(0,1,0) = Random Walk

> ARIMA(0,1,1) = Simple exponential smoothing

> ARIMA(2,0,1) = 2nd order AR, 1st order MA

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t + b_1 \varepsilon_{t-1}$$

> ARIMA(1,1,0) = 1st order AR, with differencing

$$\Delta y_t = a_1 \Delta y_{t-1} + \varepsilon_t \qquad \Delta y_t = y_t - y_{t-1}$$

> ARIMA(2,1,0) = 2nd order AR, with differencing

$$\Delta y_t = a_1 \Delta y_{t-1} + a_2 \Delta y_{t-2} + \varepsilon_t \qquad \Delta y_t = y_t - y_{t-1}$$

ARIMA + Seasonal

- Add in seasonal (cyclic) factors
- If Arima models have three factors, PDQ, then seasonal Arima models have 6: the same PDQ, and seasonal PDQ

$$Arima(p, d, q)X(P, D, Q)$$

- p = Autoregressive order (non – seasonal)
- d = Integrative part (non – seasonal)
- q = Moving Average order (non – seasonal)
- Seasonal (cyclic) parameters (lagged by a time difference)
- P = Autoregressive order (seasonal)
- D = Integrative part (seasonal)
- Q = Moving Average order (non – seasonal)

ARIMA + Seasonal

- Some examples:
- ARIMA(1,0,0) + (0,0,0) = 1st order autoregressive
- ARIMA(0,0,1) + (0,0,0) = 1st order moving average
- ARIMA(0,1,1) + (0,0,0) = simple exponential smoothing
- ARIMA(1,0,0) + (0,0,1) = 1st order autoregressive + seasonal moving average (seasonal smoothing)
- ARIMA(0,0,1) + (1,0,1) = 1st order moving average + seasonal moving average and dependence on prior season.
- ARIMA(0,1,1) + (0,1,0) = simple exponential smoothing + seasonal integrated differences.
- R demo

Time Series Using Linear Models

- We can approximate time series using linear models if we are careful.
- We can insert factors into our linear model that account for time.
  - Number of days/weeks/months/years since start.
  - Day/week/month/season of year
  - Expected highs and lows of cycle
- If our neighboring points are still related we can add in our auto-regressive terms:
  - Add a ‘time before’ and/or ‘2 times before’ values.

- R-demo

Spatial Statistics

- Spatial Statistics is a more ‘recent’ area of mathematics.
- The first comprehensive spatial statistics book was written in 1991, by Noel Cressie, “Statistics for Spatial Data”.
- Spatial data, like time series, is data that is ‘regionalized’. By regionalized we mean that the data is related to each other along multiple dimensions.
  - Time Series data is related along one dimension (time).
  - Spatial data is related along one or more dimensions.
- General spatial data is usually denoted by:

$$Y(s): s \in R$$
$$Y(s), \text{such that } s \text{ is in } R$$

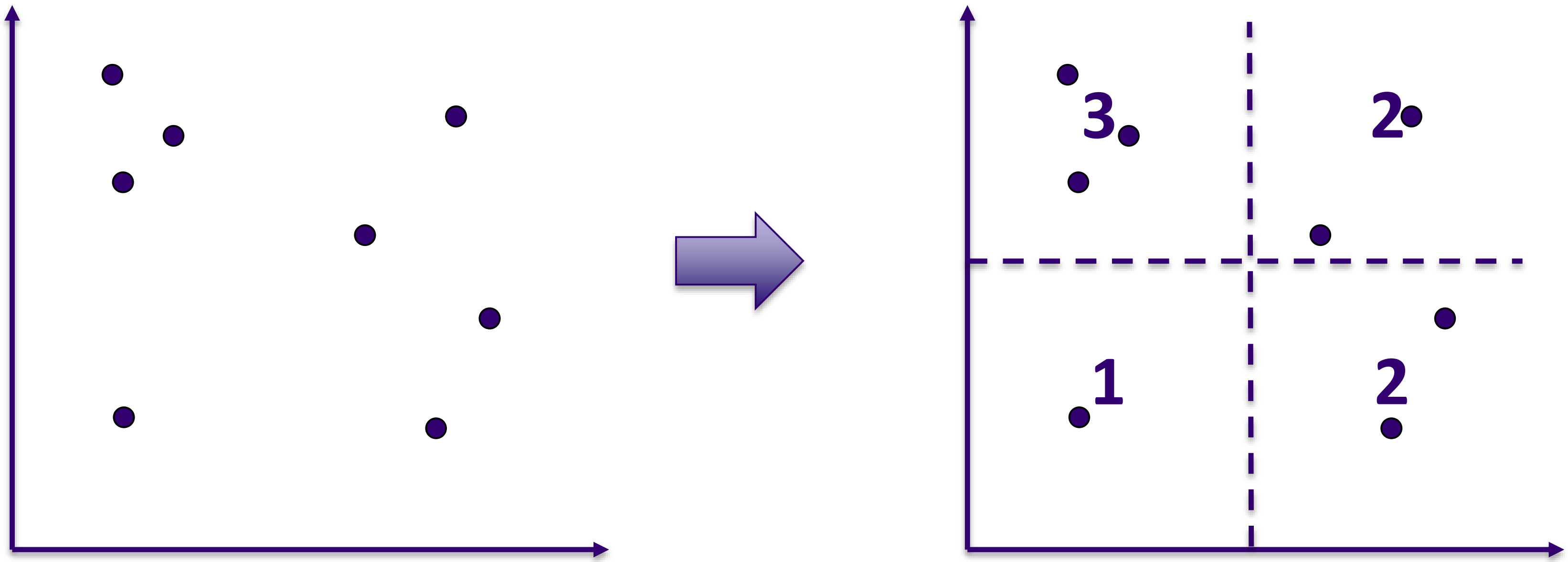
- Here, Y is our response, and s is a position vector which resides in a region, R.

Spatial Statistics

- Types of spatial data:
- Continuous Data:
  - Y is a random variable at each of the infinite continuous locations in R.
  - E.g.: Temperature, rainfall, …
- Lattice Data:
  - R is fixed, $R = \{s_1, s_2, s_3, \ldots\}$, and on a regular lattice or grid on the plane.
  - Y(s) is a random variable at these locations.
  - E.g.: aggregated measurements over an area, pixels on an image, …
- Point Process data:
  - R is fixed, $R = \{s_1, s_2, s_3, \ldots\}$, and is composed of arbitrary points on the plane.
  - Y(s) is a random variable at these locations.
  - E.g.: Mining data, most observational studies

Spatial Statistics

> We can usually format spatial data into any of these types data sets.



> Transforming into continuous data can be done via predictions or forecasting (this is called interpolation).

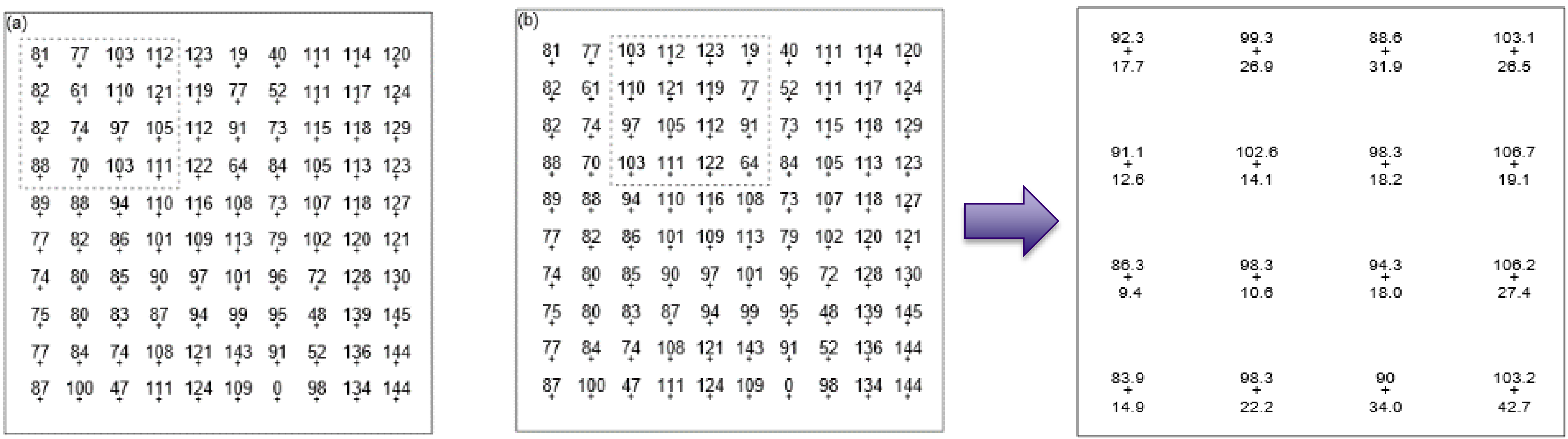
Median Polish (or mean polish)

> Spatial gridded data sets can be normalized across the axes.
> The purpose of this is to remove any linear trends in the data.
> The algorithm is as follows:
  – Take the median of each row and then subtract the median from each point in that row.
  – Compute the median of the row medians, call this the grand row median. Subtract this grand row median from each of the row medians.
  – Take the median of each column and then subtract the median from each point in that column.
  – Compute the median of the column medians, call this the grand column median. Subtract this grand column median from each of the column medians.
  – Repeat all these steps until there is no change in either of the grand medians.
> Note that using the mean for this would result in an algorithm with outlier sensitivity
> R demo

Moving Window Averages

> Just like times series, we can create a window and average across it in multiple directions.

> E.g. A grid of 10X10 points below was averaged across a sliding window of size 4X4 with overlap 2. This results in a 4X4 matrix.



* Note that this idea is used in convolutional neural networks when dealing with images.

Estimation

- Suppose you are sampling the ground for gold content at *n* locations. You then observe the resulting outcomes as *Y*.
- You might be interested in:
  - The total or average gold content across the whole region. (Global Estimation)
  - Predicting the gold content at a specific location. (Point Estimation)
- Although the methods are very similar, the point estimation accounts directly for the distances of separation between points.
  - E.g. if we wanted to predict the gold content at a specific point, reporting the average of the whole region is a poor estimate.

Weighted Averages

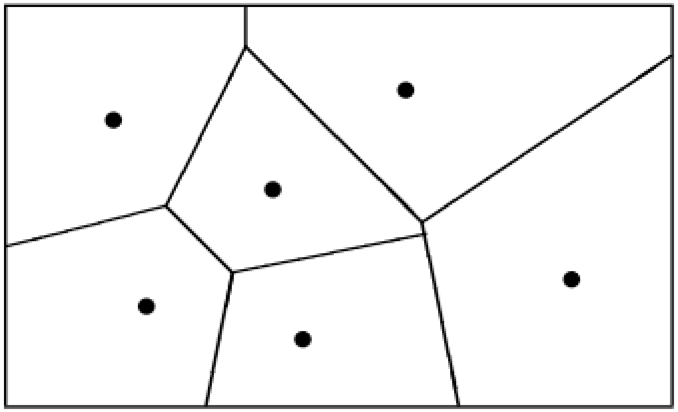
- For global estimation, you might weight each point in the average by how close it is to other points. A point separated by larger distances should have more weight than points right next to each other (which then just carry the same information).
- For point estimation, more weight is given to sites which are closer to the prediction site.

- Weighted Average formula:

$$Weighted\ Avg. = \sum w_i Y(s_i) \quad \text{where} \quad \sum w_i = 1$$

Voronoi Diagrams

> Voronoi diagrams, or polygons, split up a plane and points by creating polygons of 'shortest distance' to a point. (Also known as Delaunay Triangulation)

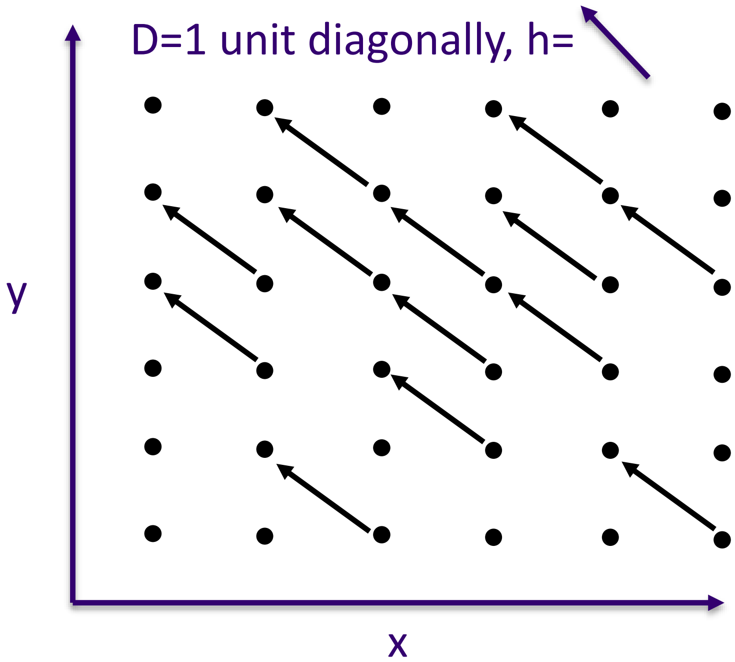


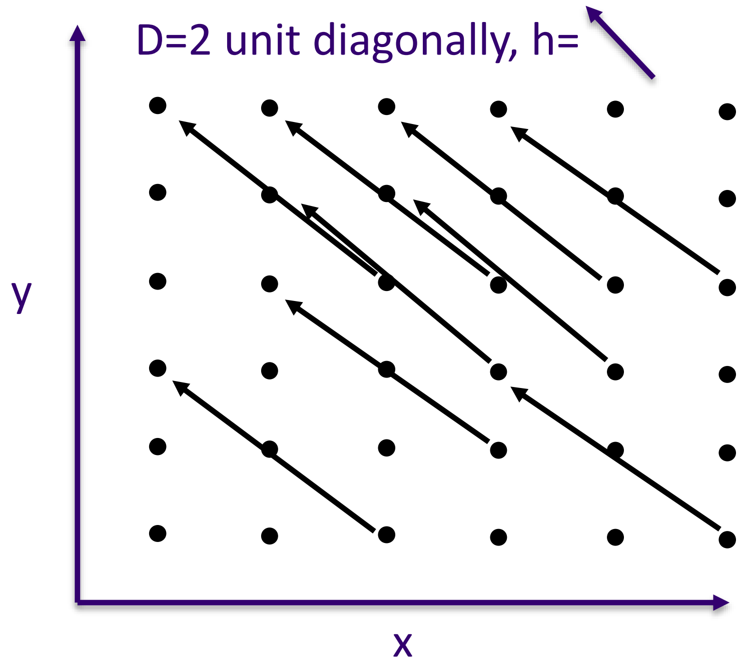
> You can imagine that we can weight the points by the area of their resulting polygons.

> R - demo

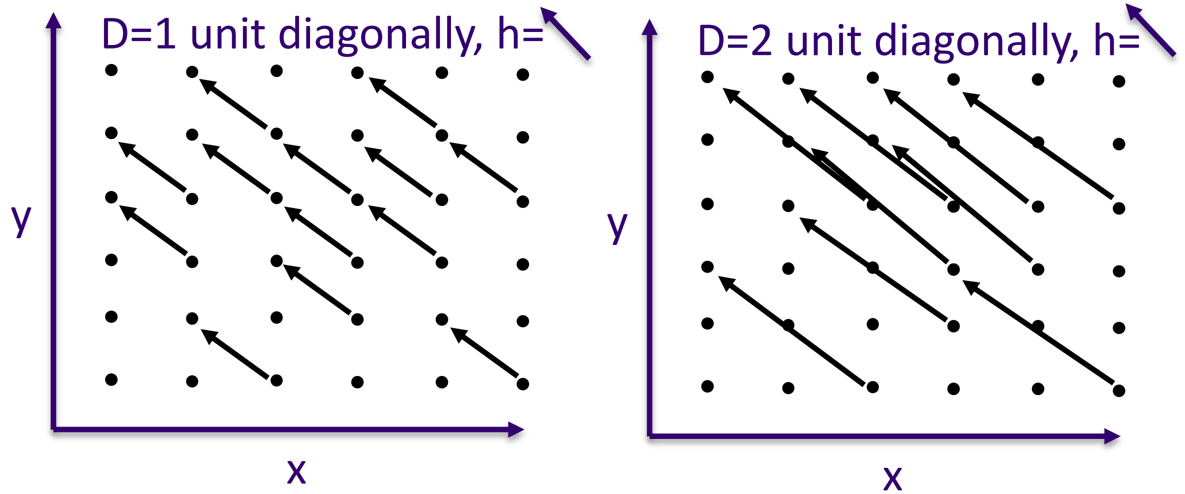
Variograms: a way to measure dependence.

- How do we measure dependence in spatial data?
- It is important to de-trend our data so that we can consider every sub region as similar regions.
  - This helps us generalize about correlations or dependence between data points.
- Consider all points separated by distance (d) in a fixed direction.

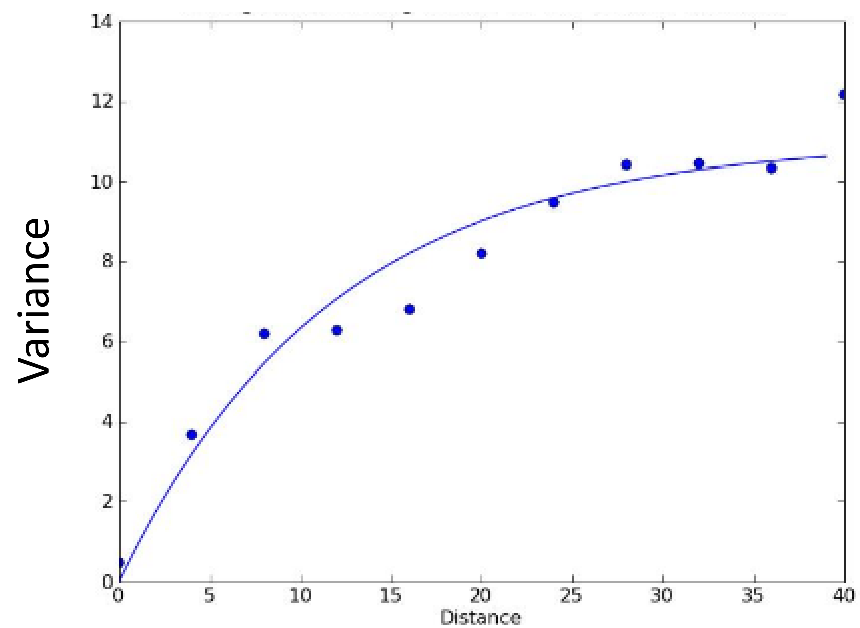




Variograms: a way to measure dependence.

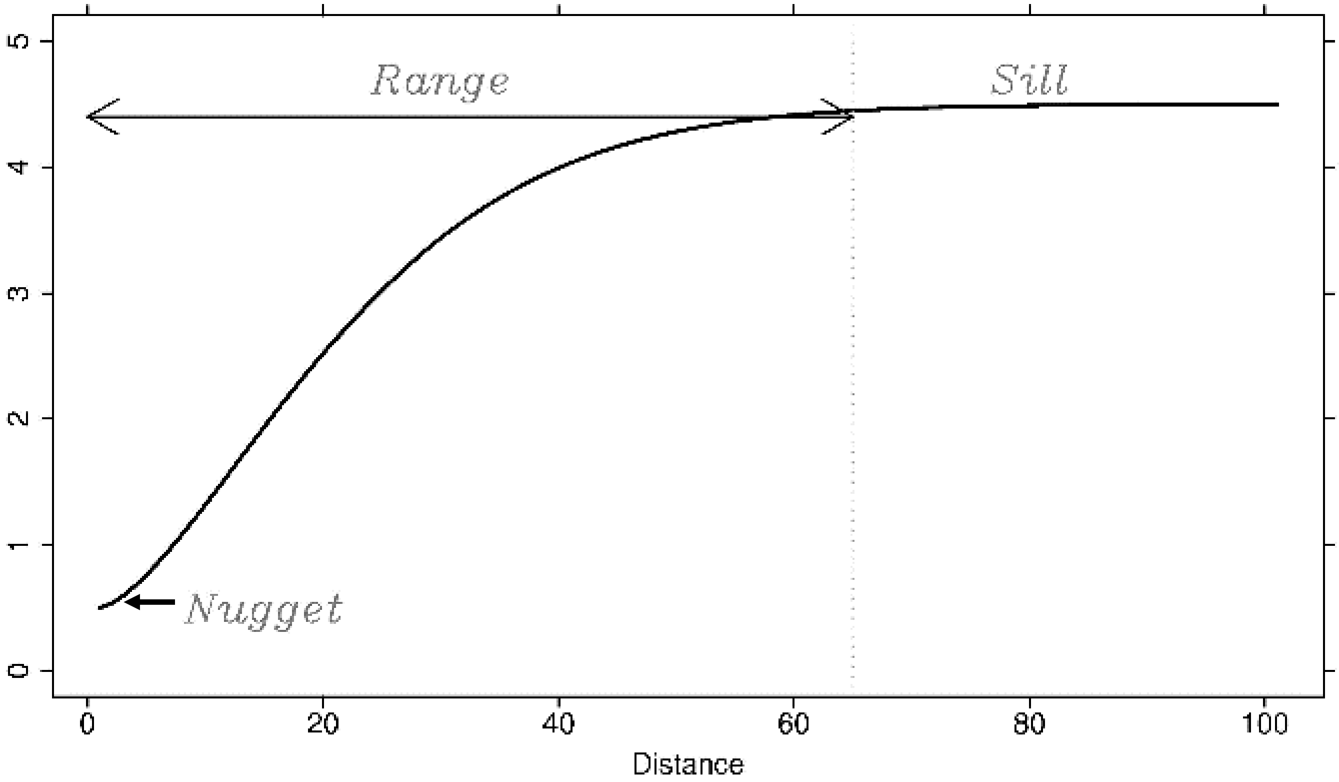


> Compute the variance of the differences between the sets as d increases:



Variograms: a way to measure dependence.

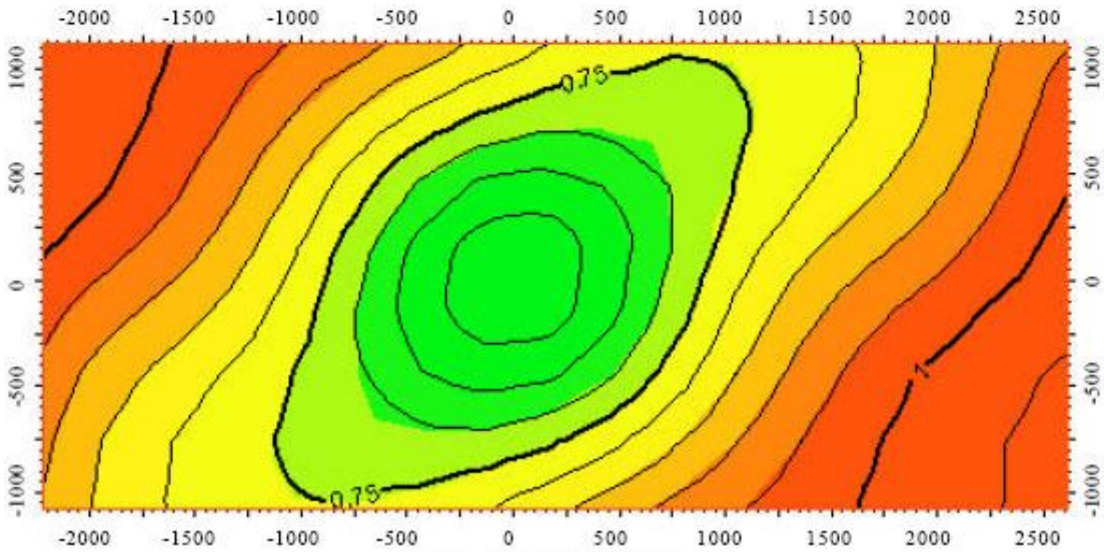
> A typical Variogram has very important properties that describe our data:



> Note, low variance in the differences implies spatial dependence.

Variograms: a way to measure dependence.

- If we plot the ranges for many directions, we end up with a ‘Rose Plot’:



- E.g., for the above plot, there is more spatial variance in the Northeast-Southwest direction than there is in the Northwest-Southeast direction.

Kriging: A word with many pronunciations.

> Most estimation procedures that we've talked about were solely based on the values and locations of points, not the relationship between the points. (how similar/dissimilar the points are)

> Kriging estimation attempts to address this issue by incorporating the variogram.

> Kriging is a form of interpolation, but weighting by the dependence in the data set (given by the variogram).

> R-demo.

Measuring Clustering

> Spatial points can be clustered, random, or overly regular.

> Human being are notoriously bad at ‘seeing’ clustering or evenness.

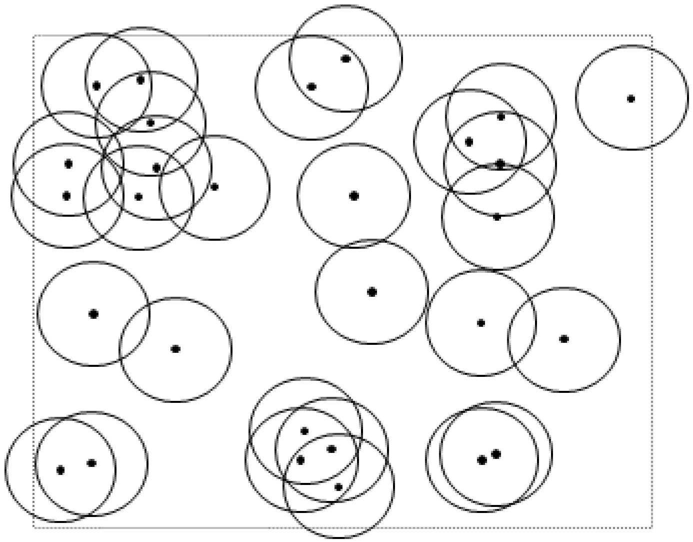
> We are interested in quantifying how spatial points cluster at different scales.

> ‘Ripley’s K’ is a common statistic used to quantify clustering.

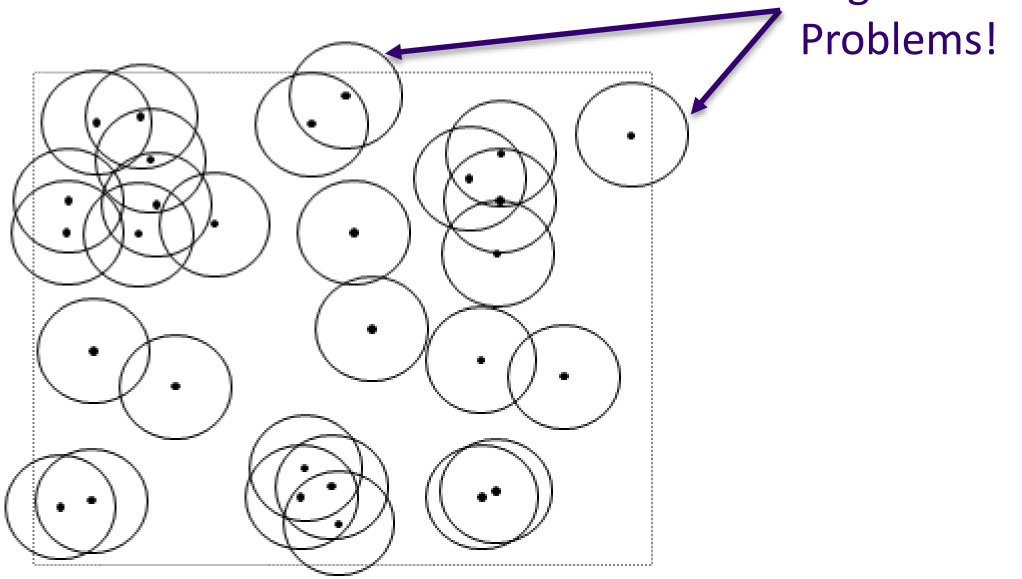
> R-demo

Ripley's K

- Computational algorithm:
  - Sample random circles at event points
  - Count how many events occur within circle of radius 'h'
  - Repeat this many times.
  - Compare this distribution to the expected distribution.



Ripley's K



- To count circles that overlap with edges, we weight the observations by a ratio of area in the region to the total area of the circle.

- R-demo

Introduction to Bayesian Statistics

- Most of the statistics we have been doing rely on assumed parameters and limiting distributions. This is called ‘Frequentist Statistics’.
- The main difference between Bayesian and Frequentist statistics is that a Bayesian view of the world includes updating/changing our beliefs when we observe data along with taking into account prior beliefs.
- Example: If we’ve lost our keys, we either
  - (1) Search our house from top to bottom.
  - (2) Search our house starting at the areas we have previously lost our keys before (laundry basket, desk, coat pockets,…), then we move onto more and more less likely places.

Introduction to Bayesian Statistics

- Using a specific way to solve some problems does not require you to sign up for a lifetime of using that exact way. In fact, the common belief is that some problems are better handled by Frequentist methods and some with Bayesian methods.

Bayes Law

- Remember the rule for conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- And

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Solving for $P(A \cap B)$

$$P(B)P(A|B) = P(A)P(B|A)$$

- Or

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

Bayes Law

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

> Applications:
  - Disease Testing: A = Have Disease, B = Tested Positive

$$P(Test+|Disease) \neq P(Disease|Test+)$$

$$P(Disease|Test+) = P(Test+|Disease)\frac{P(Disease)}{P(Test+)}$$

High Probability, usually the reported accuracy of test.

If the disease is rare, the P(disease) will be very small.

> Example:

$$P(Disease|Test+) = (0.999)\frac{0.00001}{0.0001} = 0.0999$$

Introduction to Bayesian Statistics

- What is the controversy?
  - Bayesian methods use priors to quantify what we know about parameters.
  - Frequentists do not quantify anything about the parameters, using p-values and confidence intervals to express the unknowns about parameters.

Assignment

- Complete Homework 7:
  - Perform a linear model on the combined jittered headcount and las vegas weather data set. (See homework start/hint on Moodle).
    - You want to create time/date features similar to the ones in the Dow Jones Example in class.
    - Description, dataset and homework hint on Moodle.
  - You should submit:
    - A R-script.