Affect

Affect is a verb referring to the act of influencing or causing a change in something else (not to be confused with **effect**)

Akaike's Information Criterion

AIC

$$AIC = 2nlog_e(\hat{\sigma}) + \ nlog_e(2\pi) + n + tr(S)$$

Akaike's Information Criterion (AIC) is measures the "goodness of fit" of a model while also taking into account the complexity of the model.

$$AIC_c = 2nlog_e(\hat{\sigma}) + nlog_e(2\pi) + n\left\{\frac{n+tr(S)}{n-2-tr(S)}\right\}$$

AIC$_c$

where… n is the sample size

$\hat{\sigma}$ (sigma) is the estimated standard deviation of the error term, which is estimated based on the maximum likelihood estimate ($\hat{\sigma}^2 = RSS/n$)

tr(S) is the trace of the hat matrix ($\hat{y}$)

- The **Corrected Akaike Information Criterion (AICc)**:
  - The AICc provides a trade-off between goodness-of-fit and the degrees of freedom.
  - The best model is the one with the lowest AICc
  - Note: AIC should not be directly compared to AICc
- **Applications of the AIC & AICc statistics in a GIS:**
  - AIC for an **Ordinary Least Squares (OLS) Regression** model with the AIC from a **Geographically Weighted Regression (GWR)** model; this tests whether the observed patterns in the GWR coefficient surfaces are meaningful or just due to chance. When the AIC of the OLS model is less than the AIC of the GWR model, there is likely extra, unjustified detail in the GWR model.
  - AICs for GWR models with different **explanatory variables**
  - AICs for GWR models with different **bandwidths**, the lowest AIC can be used to select the most appropriate bandwidth because AICs determine which set of surfaces result in the model that is the closest to reality.
  - AIC is often preferred over the **Cross-Validation (CV)** statistic because AICs can be used in **Poisson GWR, Logistic GWR,** and **Linear Regression** models
  - AIC is also preferred over CV because AIC takes degrees of freedom of each model into account so that they can be more accurately compared with one another

Analysis of Variance

ANOVA

Analysis of Variance (ANOVA) is a least squares analysis of qualitative data that fits means and variation about those means

- ANOVAs take separate & independent samples (each with their own mean) & gives the likelihood that they (the samples) came from the same or different populations
- ANOVAs test means by turning them into variances: One pooled variance from within the groups ("pooled within variance"), & one from between or among the groups ("variance among groups")
- R. A. Fisher's calculations are based on the marginal totals or means, averaging or summing over all observations in the treatment.
- All ANOVAs contain an error term that is a Random component of variation.
- ANOVA is appropriate if all the predictors are either qualitative or classified into a small number of groups

- **Common applications of ANOVA in a GIS:**
  - ANOVA methods are frequently used to analyze the significance of different regression models while assuming the errors are normally distributed
  - In Geographically Weighted Regression (GWR) analyses, an ANOVA can be used to test the null hypothesis that the GWR (local) model is no better fit for the phenomena being studied than the global model (ie. the **Ordinary Least Squares (OLS)** Regression model).

Terms **ABBR** **Definition**

Apoptosis

The process of tissue or cells death.

- Specifically **apoptosis** is the programmed, deliberate, or otherwise planned death of cells and/or tissue.
- Not to be confused with **necrosis**.

Atrophy

The decrease or wasting away, of an organ, tissue, or other part.

- can be the result of not using a body part enough to sustain the organ's size, an injury, or even **disease**

Attribute

Nonspatial information about a geographic **feature** in a GIS

Bandwidth

Spatial Filter Radius

h
b

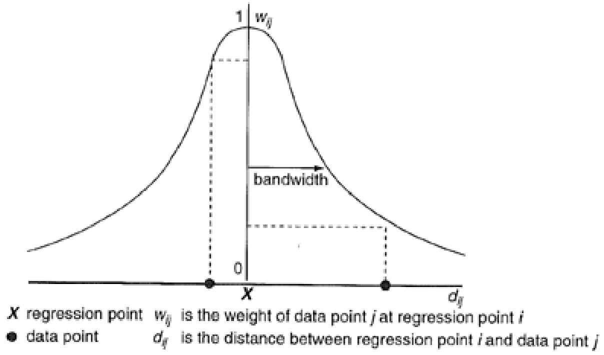


Figure 2.10 A spatial kernel (Fotheringham et al. 2002, p. 44)

$$h_{max} \approx s\sqrt[5]{\frac{243\int g(x)^2dx}{35n}}$$; for a normal $g(x)$: $$h_{max} \approx \frac{1.144s}{\sqrt[5]{n}}$$

for kernels with normal probability estimates…

$$h_{opt} = \left[\frac{2}{3n}\right]^{\frac{1}{4}} \sigma$$

where… g is the probability distribution function with a mean =0 and a variance = 1

s is the standard deviation of the sample

n is the sample size

σ is the standard deviation of the true probability estimate (standard distance can be used here)

Bandwidth (**h** or **b**) is the size of the radius of spatial analytical tools applied to individual data points.

- Essentially bandwidth defines the area around each data point in which the respective spatial statistic will be applied.

- **Bandwidth with regard to Spatial Kernels**
  - Bandwidth controls the spread of the kernel (K) hump
  - The size of the bandwidth controls the amount of **smoothing**, with increased bandwidths causing more smoothing.
    - └ Too large a bandwidth results in an over-smoothed model in which any local variation in the data is lost.
    - └ Too small a bandwidth results in an under-smoothed model, in which there is so much local variation that any larger spatial trends are hard to distinguish.
  - Bandwidth selection is an important part of any **Geographically Weighted Regression (GWR)** analysis
  - Bandwidth is a measure of the distance-decay in the weighting function of GWR using **Fixed Spatial Kernels**
  - While, the weighting function tends to have little effect on the results of a GWR model, the GWR model is highly sensitive to the bandwidth of the weighting function.

- **Bandwidth estimation methods:**
  - Terrell's **Maximal smoothing bandwidth (h_{max})**
  - Bowman & Azzelini's **Optimized bandwidth (h_{opt})**

- **Bandwidth Selection Criteria:**
  - minimization of the **Akaike Information Criterion (AIC)**
  - **Corrected Akaike Information Criterion (AICc)**
  - **Bayesian Information Criterion (BIC)**
  - **Cross-validation (CV)** minimization
  - **Least Squares Criterion**

Bayesian Information Criterion

Schwartz Information Criterion

BIC
SIC, **SBC**, or **SBIC**

$$BIC = -2log_e(L) + klog_e(n)$$

or

$$BIC = -2ln(L) + kln(n)$$

where… L is the model likelihood

k is the number of parameters

n is the sample size

Bayesian Information Criterion (BIC), also known as the **Schwartz Information Criterion (SIC, SBC**, or **SBIC)** is a statistical measure used for selecting the best models in **Maximum likelihood-based** models (such as **Regression**)

- BIC puts more weight on the number of parameters than similar measures such as **Akaike's Information Criterion (AIC)**
- BIC is not appropriate for use with large sample sizes because tends to identify models with fewer parameters then optimal

Chi-Square Statistic

Chi-Square Statistics measure the goodness-of-fit of models

- Compares categorized data with a multinominal model that predicts the relative frequency of outcomes in each category in order to estimate the extent to which they agree

Cluster

Clusters refer to areas of elevated occurrence of a specified characteristic, such that it is unlikely this characteristic could have occurred by chance alone

- **Hot Spots**: intense clustering of high values (high + Z-scores)
- **Cold Spots**: intense clustering of low values (low – Z-scores)

- **Cluster Detection and Analysis**

- **Cluster Detection and Analysis**
  - Detection and analysis of clusters can be done at 2 scales

Global Scale

- refers to whether or not clustering is present, but does not indicate where the clustering is occurring
- **Global Scale** Cluster Detection & Analysis Programs
  └ **Getis-Ord General G**
  └ **Moran's I**
  └ **Ripley's K Function**

Local Scale

- detects the presence of clustering and the specific geographic location of that cluster.
- In cases where the aetiology is unclear, the detection of specific clusters can help to identify the causal agents or modes of transmission of a disease.
- Specific clusters can also indicate areas where non-diseased individuals are likely at a higher risk of becoming diseased.
- **Local scale** Cluster Detection & Analysis Programs
  └ **Anselin's Local Moran's I**
  └ **Disease Mapping and Analysis Program (DMAP)**
  └ **Geographical Analysis Machine (GAM)**
  └ **Getis-Ord Gi***
  └ **Spatial and Space-Time Scan Statistics (SaTScan)**

Collinearity

Collinearity is when variables are **collinear** (highly correlated)

- **Local Collinearity**
- **Mulicollinearity**

- **Local Collinearity**
  - can prevent proper calibration of the spatial parameters (optimal distance or number of neighbors) for the **Geographically Weighted Regression (GWR)** based on **Akaike Information Criterion (AIC)** or **Cross-Validation (CV)** Bandwidth estimation methods.
  - Local Collinearity is indicated by Condition Numbers < 0, > 30, or = null hypothesis.

- **Multicollinearity**
  - Occurs when the independent variables are highly correlated
  - Multicollinearity is indicated by **correlation** values > 0.9, to confirm check the **Variance Inflation Factor (VIF)** values
  - Violates the assumption of independence
  - This can be devastating to the output is not accounted for
  - high correlation means that small changes in the data results in large fluctuations in the regression coefficients, which in turn causes inflation in the variance estimates & essentially makes the regression coefficients useless

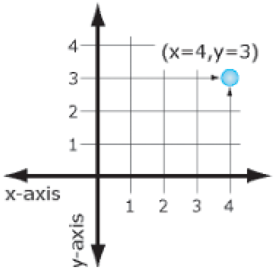
Condition Number

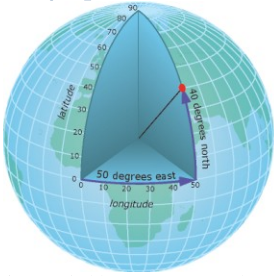
The **Condition Number** in the output of a **Geographically Weighted Regression (GWR)** analysis indicates where there may be instability as the result of local multicollinearity.

- **Local Collinearity** is indicated by Condition Numbers < 0, > 30, or = null hypothesis.

Coordinate Systems

- **Cartesian** C.S.
- **Geographic** C.S.



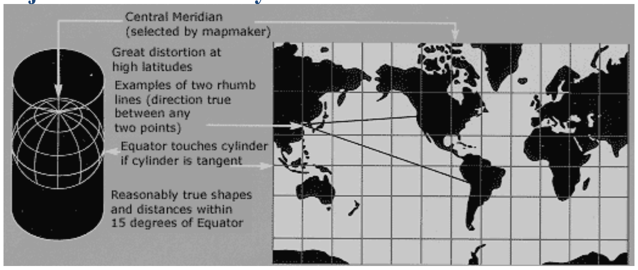


(http://webhelp.esri.com/arcgisdesktop/9.2/body.cfm?tocVisable=1&ID=24&TopicName =Georeferencing%20and%20coordinate%20systems)

A reference system used to define the exact geographic location of point on the earth's surface.

- **Cartesian Coordinate system:**
  - **x,y** coordinates define locations on a 2D, flat (planar) surface
  - x measures the horizontal, & y measures the vertical distance
  - Because the coordinate system is 2D, measures of distance, area, & direction are constant throughout the plane
- **Geographic Coordinate System:**
  - **Latitude** & **Longitude** coordinated define locations on a 3D, spherical surface
    - └ **North American Datum (NAD) 1927, 1983**
    - └ **World Geodetic System (WGS) 1972, 1984**

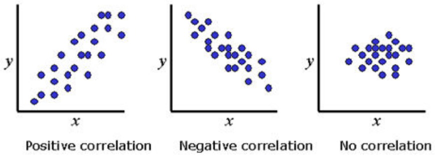
- **Projected Coordinate System**



- **Projected Coordinated System:**
  - When geographic data defined in a Geographic Coordinate System has been transformed so that it can be displayed ("projected") on a flat surface
  - This is generally done as the final step in a GIS analysis during the creation of Maps
    - └ **Universal Transverse Mercator (UTM)** zones

(http://egsc.usgs.gov/isb/pubs/MapProjections/projections.html)

Correlation



A measure of the relationship between 2 or more variables.

- **Correlation coefficient**: values range from -1.00 to + 1.00

$$S_{XY}/\sqrt{S_{XX}S_{YY}}$$ where S_{XY}, S_{XX}, & S_{YY} are the corrected crossproducts

(http://www.terraseer.com/help/stis/interface/Views/correlation_example.jpg)

Covariance

$$S_{XY}/(n-1)$$

Statistical measure of how two variables vary together

- Covariance ≠ Variance
  - └ **Variance** measures of how much one variable varies

Cross-Validation

Jackknifing

CV

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2$$

where… $\hat{y}_{\neq i}(b)$ is the fitted value of y_i

b is the bandwidth

observations for point i have been omitted from the calibration process

$$GCV = n\sum_{i=1}^{n} [y_i - \hat{y}_i(b)]^2 / (n - v_1)^2$$

GCV

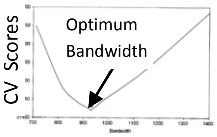
Where v_1 is the effective number of parameters in the model

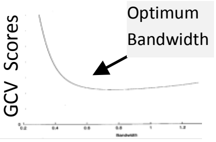
Cross-Validation (CV) is a procedure used to test the quality of the predicted data distribution by removing a data point with a known value and then using the rest of the data to predict the value for the removed point.

- CV is only possible when the regression point locations are the same as the data point locations
- Changes in **bandwidth** causes the **degrees of freedom** of the model to change, thus when CV is used to determine optimal bandwidth each of the models it compares the bandwidths of will have different degrees of freedom

- **The generalized cross-validation criterion (GCV):**
  - GCV is an approximation of the cross-validation statistic.
  - The GCV is often used instead of the CV because it is easier to calculate.
  - The v_1 term prevents the **wrap-around effect** by approaching n as the denominator approaches 0.

- **Bandwidth selection**





- **Bandwidth selection**
  - Plotting CV or GCV scores against bandwidths can indicate the most appropriate bandwidth value for a given dataset

(Fotheringham et al. 2002, Fig. 2.20, p. 60) (Fotheringham et al. 2000, Fig. 7.8, p. 181)

Terms **ABBR** **Definition**

Degrees of Freedom

df or γ

$df = n - \#of\ parameters\ being\ estimated$

For variance calculations… ***df = n-2***

- because there are only 2 parameters (β_0 & β_1)

Density

Spatial density is the number of discrete-objects per unit area

- **Simple Density**
  - the sum of points or lines that are within a given search area divided by the size of the search area
  - the result is cell density values
- **Kernel Density**
  - The values of points within a given search area are distributed across the radius of this area, such that the greatest density is closest to the location of the point and the density at the boundary of the search area is zero.
  - The sum of any distributions that intersect with one another are calculated to produce the density value for the cell
  - The results is a smooth distribution of the point densities

Dependent Variable

Observed Case (SaTScan)
Regressand
Response Variable

Y

The **Dependent (Y) Variable** is the variable or process of interest

- Regression models are used to try to predict dependent variables, by first calibrating the model with known (observed) values of the dependent variable, while **independent (X) variables** are used to better explain it.

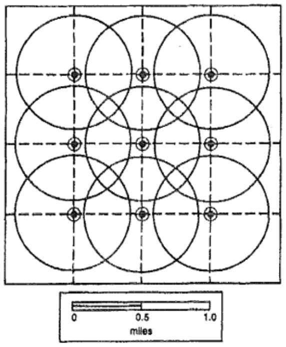
Disease

Any deviation from an organism's normal, or "healthy," state.

- This includes impairment of vital functions, organs, or systems, including interruptions, cessation, proliferation, or malfunctions, originating from abiotic and/or biotic sources
- Diseases are often diagnosed through the onset of **signs** (visual indication of harm or stress within an organism) and/or **symptoms** (non-visual, internal indication of harm or stress within an organism).

Disease Mapping & Analysis Program

DMAP



"The Regular lattice grid and the spatial filter areas to measure birth rates in the study area"
Rushton & Lolonis 1996; p. 721, Figure 3

DMAP is a spatial analysis program which performs both **Cluster Detection** and **Cluster Analysis**

- DMAP was developed by the University of Iowa's Department of Geography to study infant mortality and identify possible clustering of infant deaths (Rushton and Lolonis 1996)
- DMAP is publically available as a free download at: http://www.uiowa.edu/~geog/health/ index11.html
- DMAP can be used to spatially analyze anything containing both numerator and denominator location data; in which the numerator is the incident or event of interest and the denominator is the underlying population in which the incident has occurred.
- This program is designed to smooth the incident-rate surfaces and then identify significant rates of incident clustering using Monte Carlo simulations
- DMAP uses a methodology very similar to that used by the **Geographical Analysis Machine (GAM)**
  - └ Both DMAP and GAM aggregate all of the point level data to a circle or "filter" centered on a grid intersection point, with the grid covering the entire study area (see figure on the left)

Distribution

Spatial Distribution: is a measure of the how much something occurs within a given area

Probability Distribution: a set of probabilities that a variable will have a specific value

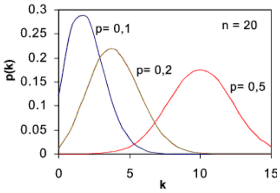
- **Binomial** Distribution

$$p(x)=\frac{n!}{(n-x)!\,x!}p^{x}q^{1-x};x=1,2,\dots n$$

where… *p* is the probability of success

q is the probability of failure (ie. *1-p*)

- **Binomial distribution:** describes the probability that exactly K successes in N independent trials, in a model designed such that the result of each trial is either success or failure.

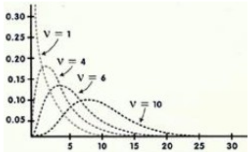


- **Discrete** distribution (***p(x)***)
- Mean = ***np***
- Variance = ***npq***

- **Chi-Square** distribution

Type equation here.

- **Chi-Square distribution:**

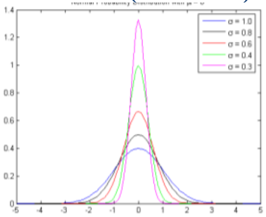


-
-

- **Normal (Gaussian)** distribution

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^{2}}$$

- **Normal (Gaussian) distribution:**

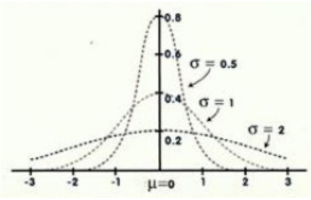


- **Continuous** distribution (***f(x)***)
- Mean = μ
- Variance = σ^2

- **Standard Normal** distribution

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}$$

- **Standard Normal distribution:**

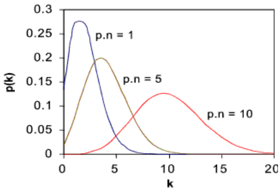


- **Normal** distribution (***f(x)***)
- Mean = ***0***
- Variance = ***1***

- **Poisson** Distribution

$$p(x)=\frac{m^{x}}{x!}e^{-m};x=1,2,\dots n$$

- **Poisson distribution:**

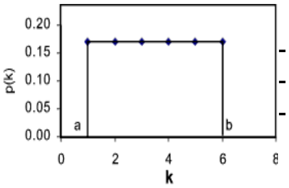


- **Discrete** distribution (***p(x)***)
- Mean = Variance = ***m*** = ***λ***

- **Uniform** Distribution

$$f(x)=\frac{1}{a};x\in[0,a]$$

- **Uniform distribution:** describes the probability that all values in the range are equally alike



- **Continuous** distribution (***f(x)***)
- Mean =$(a/2)$
- Variance =$(a^{2}/12)$

Effect

Effect is a noun, referring to the result of a change brought about by a stimulus (not to be confused with **affect**)

Endemic

When a common **disease** or disorder occurs at constant rates affecting highly percentages of the population.

- For example in Africa malaria is an *endemic* disease since there are places in which the human population is expected to get the disease at least once in their lifetime.

Environmental Systems Research Institute — **ESRI**

One of the most well known GIS software companies.
- Founded in 1969
- Products include ArcView and ArcGIS software lines

Epidemic

When a **disease affects** an abnormally high number of humans within a population, community, or region during the same period of time.
- Classifying outbreaks as *epidemics* is often subjective, as it depends on a preconceived notation of what normal infection levels would be.

Epidemiology

Study of **epidemics**, focusing on **disease** distribution, **incidence**, modes of spread, and possible methods of containment.

Within this field there are many sub-categories, such as:
- **Classical or Descriptive epidemiology**: studying populations
- **Clinical epidemiology**: studying individuals
- **Analytic epidemiology**: conducting studies to test theories

Epizootic

Epidemics which **affect** animals, non-human populations, though the **disease** may spread to the human population

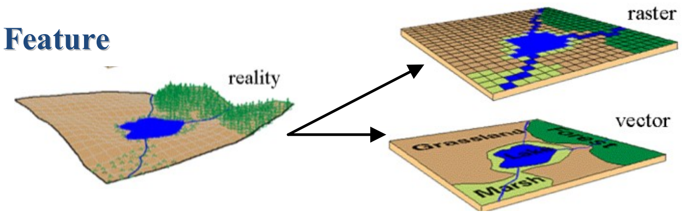
Etiology

The study of causes
- In medical fields this refers to the study of the origins of disorders, **diseases**, or otherwise abnormal conditions.

F-Statistic — **F-Stat**

A ratio of variances calculated from a sub-set of the data in order to provide information about the entire dataset.

Feature



http://oldlearn.lincoln.ac.nz/gis/gis/Intro%20to%20GIS/Intro_data_structures_test.htm

The representation of an object on a Map.
- Features must contain information defining their geographic location and their geometry
- In a GIS, features can be represented in a **Raster** data format (as cells within a grid), or in a **Vector** data format (as points, lines, or polygons).
- Vector-based features often have associated **attribute** data

Fitted Values

Estimated values (OLS, GWR)
Expected Cases (SaTScan)
Predicted values

Predicted values for the dependent (Y) variable.

Geographic Positioning System — **GPS**

A global navigation system in which handheld units receive the location data (latitude, longitude, and altitude) for their current position from Satellites orbiting Earth.

Geographical Analysis Machine — **GAM**

GAM is a cluster detection and analysis software used to detect the locations and strength of spatial clusters within point data.
- GAM uses a spatial analysis method very similar to the one used in the **Disease Mapping and Analysis Program (DMAP)**

Geographical Information System — **GIS**

A computer software-based system that can be used to analyze, capture, create, manage, present, and store spatial data and information.
- GIS systems include, but are not limited to, mapping software

Geographically Weighted Regression

$$y(u,v) = \beta_0(u,v) + \beta_1(u,v)x_1 + \varepsilon(u,v)$$

where… y is the dependent variable

x is the independent variable

u,v are the coordinates of the data

$\beta_{\#}$ are the parameters being estimated

ε is the random error term

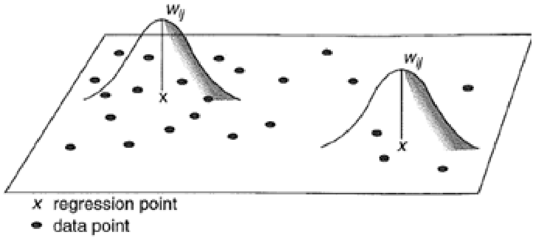


Figure 2.11 GWR with fixed spatial kernels
(Fotheringham et al. 2002, p. 45)

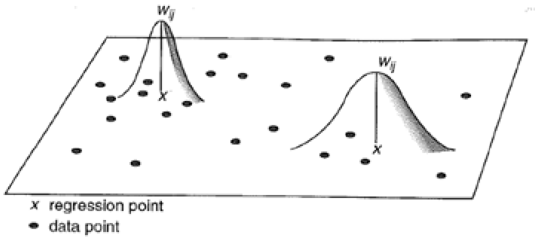


Figure 2.13 GWR with adaptive spatial kernels
(Fotheringham et al. 2002, p. 47)

GWR

Geographically Weighted Regression (GWR) fits a regression equation to every feature in the dataset, thus providing a **local model** of the variable or process being studied.

- GWR estimates the localized **Residual** values for the locations of the **Dependent (Y) variable**
- GWR allows for the estimation of the localized parameters for any point in space, not just where data was collected

- **Logistic GWR** model: response variable = 0 or 1
- **Gaussian GWR** model: the response variable could = ± ∞
- **Poisson** GWR model: the response variable = + integer counts
- **GWR with Fixed Spatial Kernels:**
  - The size of the bandwidth is pre-defined, while the number of data points that fall within this bandwidth will vary across the study area
  - Use this when the data points are regularly spaced
- **GWR with Adaptive Spatial Kernels:**
  - The number of data points that fall within this bandwidth is pre-defined, while the size of the bandwidth will vary across the study area
  - Use this when the data points are not regularly spaced , but rather clustered within the study area
  - Adaptive kernels deal with this irregularity by changing the size of the bandwidth according to the data density
  - the bandwidth size is increased where data density is low, because the data is sparser and widely distributed
  - the bandwidth size is decreased where data density is high, because the data is clustering spatially

Goodness of Fit

Degree to which a model correctly predicts the observed data

Health

The state of an organism, or part of an organism, when it is functioning optimally or at least properly, without evidence of **disease** or other malfunctions.

Histology

Microscopic Anatomy

The microscopic study of organismal anatomy

- derived from the Greek words "histo-" and "logos" translated as "a treatise of tissues."
- As its derivation implies, histology focuses on the study of the tissues of an organism and their relationship with their surrounding cellular environment.
- Histology differs from "gross anatomy" in scale – gross anatomy can be studied with the naked eye while histology can only be studied microscopically

- **Histopathology**

- **Histopathology:**
  - A branch of **pathology**, focused on the tissue changes associated with the diseased state of an organism.
  - Note histopathology can not give quantitative information on exposure, temporal changes, or fully identify parasites or pathogens.

Incidence

Total number of new **disease** cases within a specified underlying population over a specific time frame

- Do not confuse with **prevalence**

Independent Variables

Predictor Variables
Regressors
Exploratory Variables (OLS,GWR)

X's

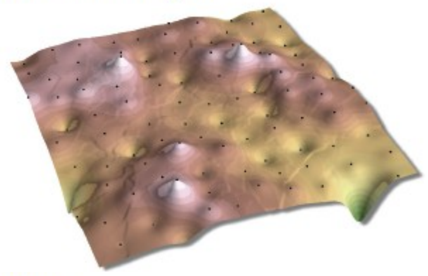
Independent (X) variables are used to model the variability in, explain the behavior of, or predict the value of the **dependent (X) variable**

Interpolation

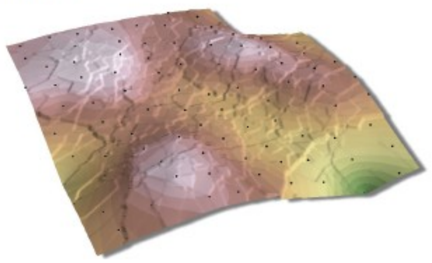
Interpolation is a type of spatial analysis in which the values of sampled data locations are used to estimate values for the surrounding un-sampled locations resulting in continuous, raster data representing the spatial nature or surface of the data

Types of Interpolation include:

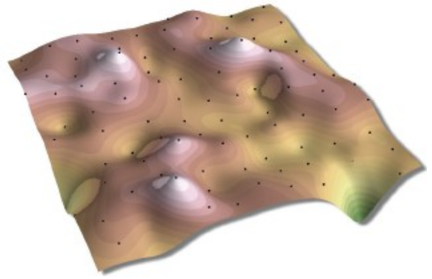
Inverse Distance Weighted



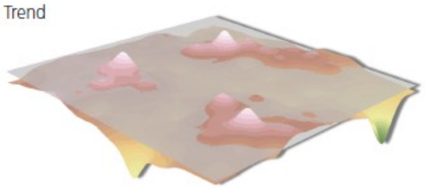
Kriging



Spline



Trend



- **Inverse Distance Weighting (IDW)**
  - An interpolation method that estimates the values of un-sampled locations by weighting them such that locations closer to sampled data points have higher weights than the un-sampled locations further away from the sampled point.
- **Kriging**
  - An interpolation method which uses geostatistical models based on **spatial autocorrelation** to weight the sampled data in order to create a prediction map of the estimated the values of un-sampled locations.
  - Weight calculations are based on the distance between the sampled data locations, un-sampled locations, and the degree of spatial autocorrelation among the sampled data
  - ○ **Co-Kriging**
    - A type of **kriging** which uses the distribution of a dataset which is highly **correlated** to the variable being studied is used along with the distribution of the primary variable in order to provide interpolation estimates
    - This can improve the accuracy of the kriging estimates when the primary dataset is small
- **Spline**
  - An interpolation method in which the values of un-sampled locations are estimated using a mathematical function which minimizes the overall surface curvature
  - The resulting interpolation has a smooth surface that passes through all of the original sampled data points
- **Trend Surface Analysis**
  - An interpolation method in which the values of the un-sampled locations are estimated by fitting a polynomial least squares regression to the sampled data points
  - The resulting interpolation minimizes the variance of the surface in terms of the input (sampled) data values
  - Unlike the spline technique, the resulting interpolation of Trend Surface Analyses rarely passes through the sampled data points.
  - This method is susceptible to outliers in the data and is thus not recommend for precise models of the spatial surface of the data; instead it is generally used to model the overall spatial "trends" of the sample data.

Jarque-Bera Statistic

- **JB Probability**

JB
JB-Prob

Kernel

- $K(x) = \frac{1}{h} g\left(\frac{x-x_i}{h}\right)$

K

Kernels are probability distribution functions, usually unimodal and symmetrical.

- Types of Kernels include:
  - **Kernel Density** estimates
  - **Adaptive distance** spatial kernels in **Geographically Weighted Regression (GWR)**
  - **Fixed distance** spatial kernels in GWR

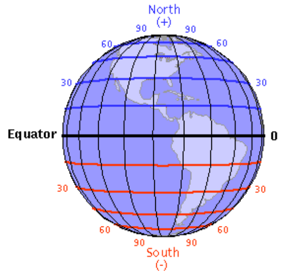
Koenker (BP) Statistic

- **K(BP) Probability**

K(BP)
K(BP)-Prob

Latitude

Parallels



Lat
y

Angular distance North (+) or South (-) of the equator.

- Lines of Latitude (“parallels”) run parallel to the equator and to each other.
- Expressed in decimal degrees; or degrees, minutes, seconds
- Used in Geographic Coordinate Systems

http://www.learner.org/jnorth/tm/mclass/Glossary_lat.html

Least Squares Criterion

$$z = \sum_{i=1}^{n} [y_i - \hat{y}_i(b)]^2$$

where $\hat{y}i(b)$ is the fitted value of y_i using a bandwidth of b

Least Squares Criterion selects the bandwidth which minimizes the value of *z*.

- This method is not an optimal bandwidth selection criterion because it tends to have the **wrap around effect**

Lesion

A rather general term referring to an abnormal region of tissue, or organ, within an organism.

- There are a number of lesion types and classifications, many of which are based on the cause or appearance of the lesion

Likelihood Ratio Test

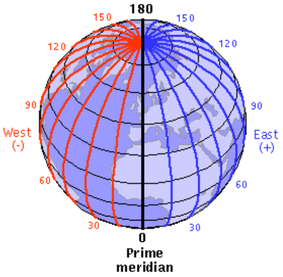
LR

The **Likelihood Ratio Test (LR)** is a test of the Logistic Model

- LR essentially tests the slope of the Logistic model against 0
- LR is very similar to the **Ordinary Least Squares (OLS)** regression model
- LR provides a **chi-squared statistic**, in which the **degrees of freedom (df)** is the difference between the two models

Longitude

Meridians

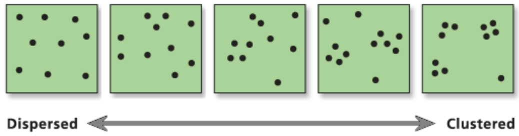


Long
x

Angular distance East (+) or West (-) of a defined meridian, usually the Greenwich Prime Meridian.

- Lines of Longitude are equally-sized circles that intersect each other at the North and South poles
- Expressed in decimal degrees; or degrees, minutes, seconds
- Used in Geographic Coordinate Systems

http://www.learner.org/jnorth/tm/mclass/Glossary_long.html

Terms	ABBR	Definition
Modifiable Area Unit Problem	**MAUP**	
Monte Carlo Simulation • **Markov Chain Monte Carlo**	**MCMC**	
Nearest Neighbor Analysis		
Necrosis		Death of a once living tissue or cells - derived from the Greek word "nekros" meaning dead body - Specifically necrosis is the unprogrammed, accidental, or otherwise unnatural death of cells and/or tissue. - Not to be confused with **apoptosis**.
Ordinary Least Squares Regression	**OLS**	**Ordinary Least Squares (OLS)** uses a single regression equation to explain the variable or process being studied, and thus provides a **global model** of the respective phenomena.
Pandemic		An **epidemic** occurring on a global scale
Panzootic		An **epizootic** occurring on a global scale.
Pathogen		A specific causative agent of **disease**. - Pathogens can be infectious **biotic** organisms (bacteria, viruses, fungi, or other microorganisms), or noninfectious **abiotic** agents (chemicals, environmental changes, etc).
Pathology		The study of **disease** - derived from the Greek words "pathos" and "logos" translated as "a treatise of disease." - Modern pathology is defined as the medical branch which studies and treats the essential nature of diseases, specifically the anatomic and/or physiological changes the disease elicits within the affected organism.
• **Pathobiology**		• **Pathobiology:** - The biological study, or practice, of **pathology**
Pathopnmonic		The specific **signs** and/or **symptoms** associated with a specific **disease** or causative agent.
Point Pattern Analysis  	**PPA**	Geographical analysis of the spatial patterns and overall nature of points, usually in the form of individual case locations. - This type of spatial analysis is especially common in biological, epidemiological, or crime-based studies - The main geospatial trends identified in PPA are: └ **Clustering**: Points spaced closer together with higher concentrations or densities, than would be expected under a normal distribution └ **Dispersion:** Points spaced further apart with lower concentrations or densities, than would be expected under a normal distribution └ **Random Distribution**: Points randomly distributed in space, following a normal distribution. Neither **clustered** nor **dispersed.**

Prevalence

Total number of existing **disease** cases over a specific time frame

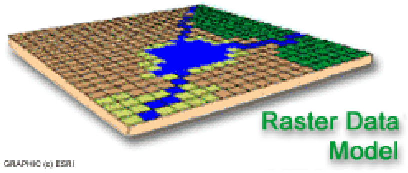
- Provides a summary of the current burden of the disease within the underlying population
- Do not confuse with **incidence**

Probability

A statistical measure of the likelihood of the occurrence of a particular outcome given a set of possible outcomes

- Probability estimates range from 0 to 1
  - └ ***P(a) = 0.0*** : completely impossible outcome
  - └ ***P(a) = 0.5*** : unpredictable outcome
  - └ ***P(a) = 1.0*** : completely certain outcome

Raster



The representation of spatial data in a **GIS** in the form of a grid of equally sized grid cells (“pixels”), each with its own value.

- Raster data is **continuous**, each data point (pixel) has a value

(http://lagic.lsu.edu/gisprimer/whatsgis.asp?topic=howitworks&sub=data)

Regression

A statistical method for measuring the relationship between a single **dependent (Y) variable** and one or more **independent (X) variables** which could be influencing the Y variable.

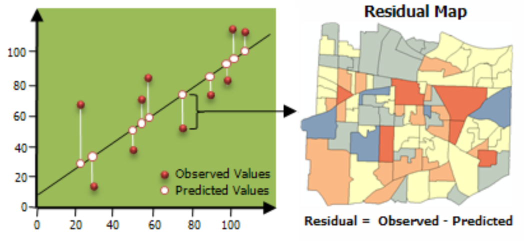
- The results of Regression analyses can be used to determine whether or not certain independent variables are actually influencing the dependent variable, and if so how muc
- Regression methods can also be used to predict the value of the dependent variable

- **Types of Regression include:**
  - **Geographically Weighted Regression (GWR)**
  - **Ordinary Least Squares (OLS)** Regression
  - **Standard Linear Regression (SLR)**

Residual

Observed/Expected Deviation

e_i
ODE



The difference between the **observed** & **expected** values of the **Dependent (Y) Variable** in **Regression** models

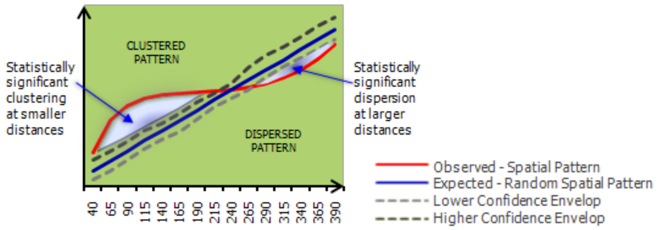
- they essentially represent the unexplained nature of Y
- they can be used to estimate the fit of the regression model
  - └ smaller residuals indicate a well fitting model
  - └ larger residuals indicate a poor fitting model
- Residuals are assumed to follow a **normal distribution**, with a mean = ***0***, and a variance of σ^2

http://webhelp.esri.com/arcgisdesktop/9.3/body.cfm?tocVisable=1&ID=1&TopicName=Residuals%20Grapic

StdResid

- **Standardized Residual:** standardized values of the Residuals
  - where the mean = 0 and the standard deviation =1.
  - StdResid > 2 indicates model under-prediction
  - StdResid < -2 indicates model over-prediction

Ripley’s K Function



Robust

Robust data is data that performs well with assumed normality

Terms	ABBR	Definition

Root Mean Square Error

RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i^*)^2}$$

A statistical measure of the difference between the locations of known or sampled data and the locations estimated by interpolation or digitizing methods

- RMSE is the standard deviation of samples from a known set of observed (or sampled) data points (x_i*)

R-squared

Coefficient of determination

R^2 or **R2**

R-squared (R^2) is the statistic calculated by **regression** analyses in order to quantify the performance of the model, in terms of explaining the variation in the **dependent (Y) variable**

- R^2 values range from 0 to 100 percent (ie. 0 to 1)
  - └ R^2 = 0.00 : the model explains none of the variation in Y
  - └ R^2 = 0.50 : the model explains 50% of the variation in Y
  - └ R^2 = 1.00 : the model explains 100% of the variation in Y

- **Adjusted R^2**

$$R^2_{adj} = \frac{(n-1)R^2 - k}{n-k-1}$$

AdjR2 or R^2_{adj}

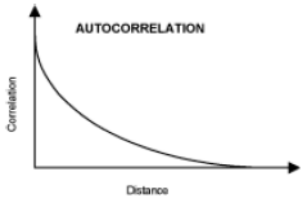
- **Adjusted R-square (R^2_{adj})**
- Accounts for the number of parameters in the model by modifying the coefficient of determination

Spatial and Space-Time Scan Statistics

SaTScan

SaTScan was developed by Martin Kulldorf to analyze spatial, temporal, and spatio-temporal data of health events using scan statistics

Spatial Autocorrelation

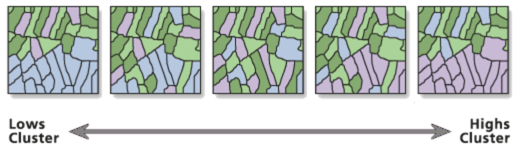


Spatial Autocorrelation is a statistical measure of the degree of spatial clustering present in the dataset based on both the values and the locations of the data points.

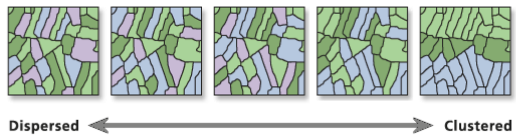
- Clustered data is indicated by Positive spatial autocorrelation
- Dispersed data is indicated by Negative spatial autocorrelation

Global Statistics

- **Getis-Ord General G**



- **Moran's I**

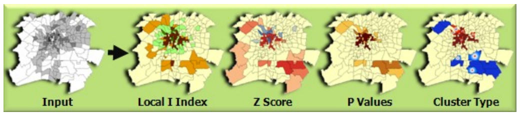


Global Measures of Spatial Autocorrelation include:

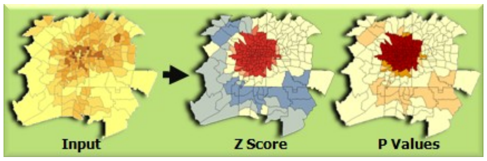
- **Getis-Ord General G**
  -
  -
  -
  -
- **Moran's I**
  -
  -
  -
  -

Local Statistics

- **Anselin's Local Moran's I**



- **Getis-Ord Gi***



Local Measures of Spatial Autocorrelation include:

- **Anselin's Local Moran's I**
  -
  -
  -
  -
- **Getis-Ord Gi***
  -
  -
  -

Spatial Epidemiology

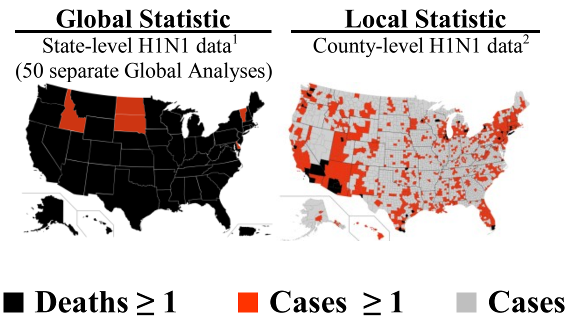
The purpose of spatial epidemiology is to first describe variations in the spatial patterning of diseases, and second to perform analyses on this data, the results of which will hopefully further our understanding of the disease

Within spatial epidemiology there are four categories of study:

1) disease mapping
2) geographical correlation studies
3) risk assessment in relation to a point or line-source
4) detection of disease clustering

Spatial Statistics

Geostatistics



Global Statistic

State-level H1N1 data[1]
(50 separate Global Analyses)

Local Statistic

County-level H1N1 data[2]

■ **Deaths ≥ 1** 🟥 **Cases ≥ 1** ▨ **Cases = 0**

[1] http://h1n1-virus.info/

[2] http://commons.wikimedia.org/wiki/File:Swine_flu_infections_and_deaths_by_county_June_2009.svg

Spatial Statistics are important in geospatial analyses because they are designed for geographic data, which by its nature tends to violate many of the assumptions of ordinary statistical analyses (ie. Normality, autocorrelation, etc.)

- **Global Spatial Statistics:** when a spatial analysis is applied at the "global" level, one set of results is produced which represents the general or average trend across the study area.
  - Few applications in a GIS environment
  - GIS Analyses based on Global Spatial Statistics include:
    - └ **Ordinary Least Squares (OLS) Regression**
    - └ **Linear Regression**
- **Local Statistics:** when a spatial analysis is applied at the "local" level, a separate set of results is produced for each location in the study sample. By mapping these local results spatial variation can be identified within the study area.
  - Many applications in a GIS environment
  - GIS Analyses based on Local Spatial Statistics include:
    - └ **Geographically Weighted Regression (GWR)**
    - └ **Point Pattern Analysis**

Standard Deviation

for distributions: $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}^2}$

StDev or **σ**

Standard deviation (StDev) is a statistical measure of the deviation of observations from their mean

Standard Distance

standard distance deviation

Standard distance is a statistical measure of the compactness of the spatial distribution of features around the estimated mean center of their distribution

- In a GIS this is usually depicted as a circle around the mean center of the data locations, in which the radius of the circle is the standard distance

Standard Error

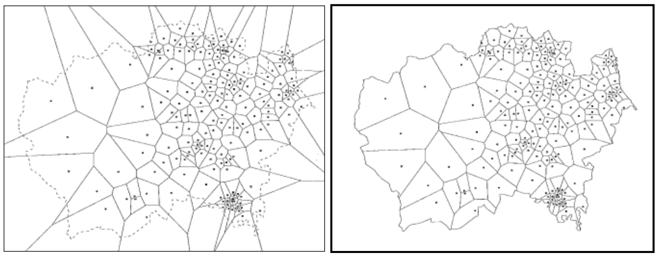
estimation of the **standard deviation** of the sampled distribution

Stressor

Something which triggers a stress response from an organism.

- Stress is a natural part of life and generally makes organisms more resilient; however, too much stress can cause an organism to deteriorate resulting in significant problems and an unhealthy (and potentially **diseased**) state of being
- Uninterrupted, prolonged, or unexpected stressors can prove to be deadly.

Thiessen Polygons



(Fotheringham et al. 2000; Fig. 3.6(a-b), p. 39-40)

Polygons generated from data point locations such that each polygon represents the area of influence of its data point

- generally created for irregularly spaced point data

Thiessen Polygons are created such that…

- each data point falls within its own polygon
- each data point is closer to the center of the polygon encompassing it than it would be to the center of any other polygon in the study area
- there are no polygons without a data point within them
- When initially created the polygons around the parameter tend to have exaggerated areas extending outside the study area. This can be fixed by "clipping" the thiessen polygons layer to the boundary of the study area (see Figures on the left).
-

Variance

Var
σ^2 or **s^2**

A statistical measure of the deviation from the mean of the values in a distribution

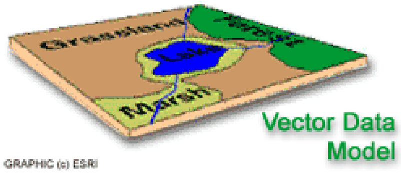
Variance Inflation Factors

VIF

Variance Inflation Factors (VIF) are a statistical index measure of the amount of variation in the coefficient has increased due to **collinearity**

- VIF can be used to detect severe **multicollinearity**

Vector



The representation of spatial data in a **GIS** in the form of points, lines, or polygons generally representing features.

- Vector data is **discrete** (i.e. the features have boundaries)

(http://lagic.lsu.edu/gisprimer/whatsgis.asp?topic=howitworks&sub=data)

Wald Statistic

- **Wald Probability**

Wald
Wald-Prob

Wrap-around effect

When the calibration wraps itself around the data points.

- This is an issue when trying to estimate the model bandwidth.

Z-score

A statistical measure of a value's deviation from the mean

- Z-scores are expressed in **Standard Deviation** units
- In a normal distribution…
  - └ 68% of the values will have Z-scores of ± 1.00, this means they are within 1 standard deviation of the mean
  - └ 95% of the values will have Z-scores of ± 1.96, this means they are within 2 standard deviations of the mean
  - └ 99% of the values will have Z-scores of ± 2.58, this means they are within 3 standard deviations of the mean
- Z-scores are often used to compare data that have different distributions, means, & standard deviations