# Correlation notes:

Spurious correlation – seem connected, but aren’t. Can be by chance or by the influence of a separate variable on both.

Autocorrelation – correlation between values of the same variable over different measurements (eg. temperature on different days.)

* A correlation coefficient measuring the correlation between a variable and itself
* Can be used to determine non-randomness of data
* To classify an appropriate time series model for random data

Cross correlation – between two timeseries for different, but corelated variables

Spatial or time

Pearson’s product moment correlation (parametric test)

Correlation coefficients estimate the strength of the relationship between x and Y. However, do not tell you the probability that the correlation occurs through random chance.

-small samples more likely to see incorrect correlations

-to test take r = coefficient of determination. How much variation in the y axis can be explained by variation in the x axis? However, this assumes that x influences y. Correlation does not equal causation.

Assumptions

* Sample pairs are independent
* X and y follow normal distributions
* Homoscedasticity
* Linearity of data
* X and Y and not confounded (x does not include a measurement of y)

# Autocorrelation

<https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/slides-autocorrelation_kriging.pdf>

Correlation of a variable with itself.

## Spatial Autocorrelation

The values of the variable area correlated across different locations. Somewhat a function of distance between them,

Positive autocorrelation: closer points more similar

Negative autocorrelation: closer points less similar

Causes:

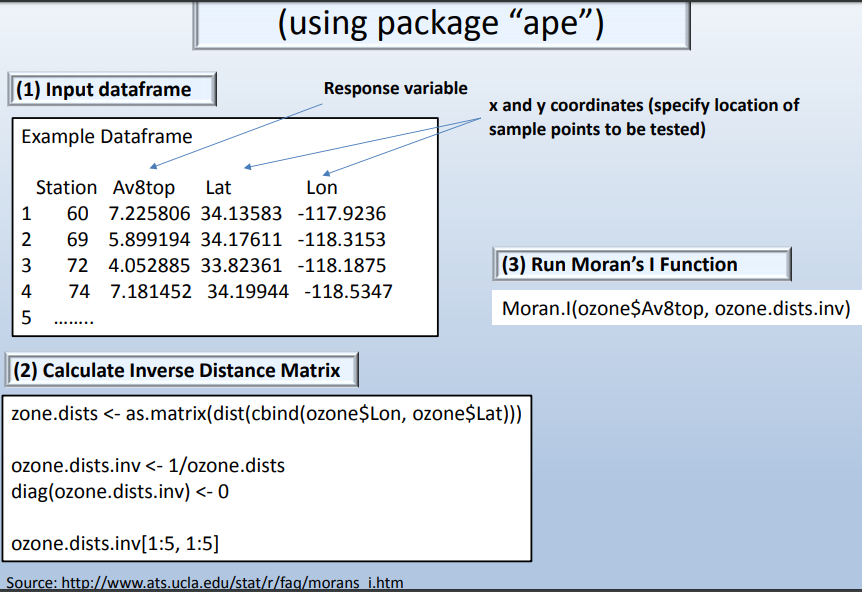
* Sample sites not being random
* Interaction of variables across space
  + Univariate: response variable is correlated with itself (eg. more plants near plants, dur to seed)
  + Multivariate: interaction of response and predicter variables. Eg. more plants in wetter area

Good as it allows for interpolation (estimation of values at non-sampled sites)

Bad as it indicates that observations are not independent. This violates the assumptions of several statistical tests and can lead to inaccurate conclusions

Testing for spatial autocorrelation

* Moran’s Index measures degree of correlation between samples based on both value and distance between points – lets you know if data is random, clustered or dispersed which will impact autocorrelation
* Measures the correlation of a variable with itself at different points in space (see graph)
* Ranges from -1 to 1. 0 = no SAC
* Null hypothesis = no autocorrelation. P < 0.05 means that autocorrelation exists
* Sensitive to extreme values
* Can also use Geary’s C,



If you want to avoid spatial autocorrelation be careful with your experimental design. Particularly spacing

There are also some statistical methods to account for autocorrelation. See Dale, M.R.T., Fortin, M. 2002 Spatial autocorrelation and statistical tests in ecology. Écoscience. 2002; 9(2):162–167.

<https://www.rdocumentation.org/packages/tseries/versions/0.1-2/topics/acf>

above focusses on temporal autocorrelation

<https://cran.r-project.org/web/packages/acnr/vignettes/autocorrelation.html>

<https://rspatial.org/raster/analysis/3-spauto.html>

Temporal autocorrelation: variables correlation with itself over time. Generally, the variable will be more correlated over shorter time scales

Can use the cor() function with vectors a and b (subsets of a larger vector) to assess autocorrelation. This is a “one-lag correlation meaning that the values are compared to their immediate neighbours and not others

Can also use the acf() function on the larger vector. This will create a graph showing the degree of autocorrelation at different lags. Generally smaller lags yield larger autocorrelation values.

## Spatial Autocorrelation

-multiple dimensions, can be harder to determine which objects are nearer to each other.

-How similar are objects at different locations to eachother. If there is a high degree of spatial autocorrelation, regions with similar values will be spatially clustered.

Causes:

* Exogenus: caused by another spatially autocorrelated variable eg. rainfall
* Endogenus: caused by a process, eg. the spread of a disease

Common statistic to describe spatial autocorrelation is Moran’s I

Need to define what is “near” as likely that measurement is not included in your data.

Moran() function

To test significance use the Monte Carlo simulation moran.mc

* Random values assigned to each observation and moran’s I computed. This is repeated several times to create a distribution of expected values
* Observed Moran’s I compared to theoretical distribution to assess how likely it is that the observed value would occur under a random draw

Moran.plot()

<https://www.statisticshowto.datasciencecentral.com/morans-i/>

Moran’s I

-a measure of spatial autocorrelation ie, how similar an object is to those surrounding it.

-1 to 1

-1 = perfect clustering of dissimilar values

0 = no autocorrelation, values perfectly random

+1 = perfect clustering of similar values

Is an inferential statistic – so needs to be tested for significance

<https://stats.idre.ucla.edu/r/faq/how-can-i-calculate-morans-i-in-r/>

install.packeges(“ape”

library(“ape”)

Need to create a matrix of inverse distance weights. Entries for pairs that are closer together are higher than for pairs further apart

First make a matrix. One way to do this is below

ozone.dists <- as.matrix(dist(cbind(ozone$Lon, ozone$Lat)))

**ozone.dists.inv <- 1/ozone.dists**

**diag(ozone.dists.inv) <- 0**

**ozone.dists.inv[1:5, 1:5]**

Then calculate Moran.I

Moran.I(ozone$Av8top, ozone.dists.inv)

Will give you observed, expected then p value to say if there is spatial autocorrelation

# Spatial autocorrelation and statistical tests in ecology

-Most parametric tests include the assumptions that observations are independent. This is not often the case in the natural world.

-eg. if there are n units of information, we don’t have an effective sample size of n because some units area not independent of each other. Effective sample size is less and you will find significant results more often than is justified by the data in tests with an assumption of independence

-Can mitigate this by experimental design or statistical methods

-First, the degree of spatial autocorrelation needs to be determined on all pairs of samples of a given distance. Ie using Moran’s I or Geary’s c. Done for a series of distances which is plotted on a spatial correlogram

-autocorrelation may effectively drop off with distance, or cycle (see figure 1)

Ways of dealing with the problem

* Subsampling
  + If you have a large sample size, can use a subset with low autocorrelation. Eg. only use samples far enough from eachother than autocorrelation is close to zero
  + Not effective if the autocorrelation does not stay at zero
* Adjusting type I error
  + Increase type I error in recognition of autocorrelation
  + There are no set rules on how much to increase the error value
* Adjusting effective sample size
  + In some cases, can determine how much the test statistic should be shrunk to account for autocorrelation
  + In tests of the mean variance of the mean is estimated as sample variance/sample size (ie mean variance = s2/n
  + Cressie 1991 suggests that mean variance can be adjusted to take autocorrelation into account by using the covariance of different samples.
  + 
  + If there is no autocorrelation, this will equal s2/n, if there is autocorrelation s2/n’. Lots of math equations to find this.
  + Can be sued for one and two sample t-tests, and ANOVA f-tests
  + Effective sample size decreases with the range and magnitude of positive spatial. Autocorrelation, negative autocorrelation can increase the effective sample size
  + Multiple runs of n’ tests can produce different numbers, which can make it unsuitable for effective data analysis
* Accounting for spatial structure – partitioning out or factoring in
* Randomization methods
  + Randomization/permutation – compare a test statistic generated by the original data with a distribution of the same statistic calculated form the data once it has been randomized in some way.
  + Restricted randomizations – spatial structure preserved

Suggests that the model procedure is the best currently available – allowing estimation of the effective sample size

Pilot study to asses autocorrelation in the area before fill study

-Larger data sets offer more flexibility in dealing with spatial autocorrelation

-Data may not fit the assumptions on which the models are based

<http://naes.unr.edu/shoemaker/teaching/NRES-746/Spatal_Autocorrelation.html>

# Data for test

|  |
| --- |
| <https://stats.idre.ucla.edu/r/faq/how-can-i-calculate-morans-i-in-r/> |
| UCLA Institute for digital research and education |
| Top ozone measurement from ten weather stations in the Los Angeles Area over ten months |

<http://geog.uoregon.edu/GeogR/topics/variograms.html>

Variogram/semi-variogram displays variance within a group of objects, plotted as a function between observations

<http://rfunctions.blogspot.com/2017/06/how-to-identify-and-remove-spatial.html>

<http://rstudio-pubs-static.s3.amazonaws.com/9687_cc323b60e5d542449563ff1142163f05.html>

<https://rstudio-pubs-static.s3.amazonaws.com/456209_0fa02e75adc243dca9a2d1e51816889f.html>

Bat data

(Dale and Fortin, 2002)

Dale, M.R.T., Fortin, M.-J., 2002. Spatial autocorrelation and statistical tests in ecology. Écoscience 9, 162–167. https://doi.org/10.1080/11956860.2002.11682702

<http://naes.unr.edu/shoemaker/teaching/NRES-746/SpatialAutocorrelation.html>