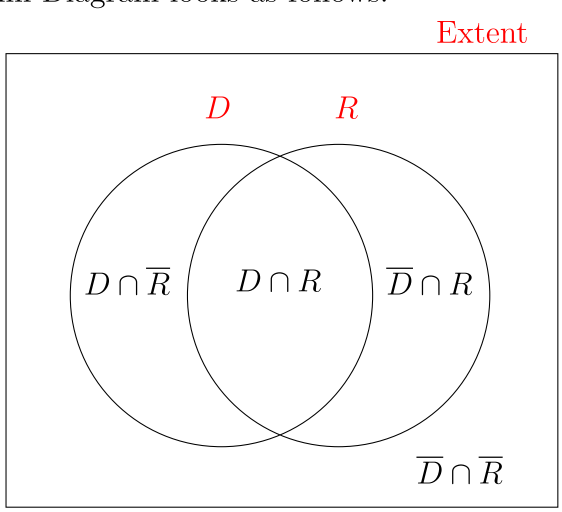
Advanced GIS - Class Notes on Weight of Evidence

Eugenio Arima

November 30, 2018

1 Example using deforestation and 1 km buffer distance to roads

We are interested in examining whether there is any evidence that deforestation is more prevalent closer to roads or not. We have two binary raster files: deforestation, represented by the letter D and within 1 km from nearest road, represented by R. The Venn Diagram looks as follows:



Cross-tabulate both maps to obtain a 2×2 table.

	R	$\overline{R}$	
D	T_{11}	T_{10}	$T_{1.}$
$\overline{D}$	T_{01}	T_{00}	$T_{0.}$
	$T_{.1}$	$T_{.0}$	$T_{..}$

The conditional probability of D occurring given the presence of R is written as $P(D|R)$. Conditional probability is defined as $P(D \cap R)/P(R)$. This can be expressed in terms of cross-tabulation areas as:

$$P(D|R) = \frac{P(D \cap R)}{P(R)} = \frac{p_{11}}{p_{.1}} = \frac{T_{11}}{T_{.1}}$$

where the little p is for probability. The last equality basically says that these probabilities can be expressed in terms of areas.

We can express the relationship between D and R in terms of conditional odds (probability of occurrence over probability of non-occurrence). Since these are binary maps, these values can be easily determined as follows:

$$O(D|R) = \frac{P(D|R)}{1 - P(D|R)} = \frac{P(D|R)}{P(\overline{D}|R)}$$

This can also be expressed in terms of area:

$$O(D|R) = \frac{p_{11}/p_{.1}}{p_{01}/p_{.1}} = \frac{p_{11}}{p_{01}} = \frac{T_{11}}{T_{01}}$$

Similarly, we can calculate the conditional odds of D given the *absence* of R, $O(D|\overline{R})$:

$$O(D|\overline{R}) = \frac{p_{10}/p_{.0}}{p_{00}/p_{.0}} = \frac{p_{10}}{p_{00}} = \frac{T_{10}}{T_{00}}$$

Combining the two conditional odds expressions we obtain a measure of association between the two binary patterns known as the **odds ratio** O_R, defined as:

$$O_R = \frac{O(D|R)}{O(D|\overline{R})} = \frac{T_{11}T_{00}}{T_{01}T_{10}}$$

If we take the natural log of this expression, we convert the odds ratio to a logit scale. This new index is called **contrast**, C_W.

$$C_W = \ln O(D|R) - \ln O(D|\overline{R})$$

These results can also be achieved by using an odds formulation, which will give a nice Bayesian interpretation of posterior probability, given information and naive probability.

From Bayes law, we have:

$$P(D|R) = \frac{P(R|D)P(D)}{P(R)}, \quad P(\overline{D}|R) = \frac{P(R|\overline{D})P(\overline{D})}{P(R)}$$

We can obtain the odds $O(D|R)$ from the above equalities:

$$O(D|R) \equiv \frac{P(D|R)}{P(\overline{D}|R)} = \frac{P(R|D)P(D)}{P(R|\overline{D})P(\overline{D})} = O(D)\frac{P(R|D)}{P(R|\overline{D})}$$

because $\frac{P(D)}{P(\overline{D})}$ is just the odds $O(D)$. The term $\frac{P(R|D)}{P(R|\overline{D})}$ is known as the *sufficiency ratio.*

Likewise, we can write:

$$O(D|\overline{R}) = O(D)\frac{P(\overline{R}|D)}{P(\overline{R}|\overline{D})}$$

and the term $\frac{P(\overline{R}|D)}{P(\overline{R}|\overline{D})}$ is called *necessity ratio.*

If we take the log of those odds, we obtain the definition of weights W:

$$\ln O(D|R) = \ln O(D) + W^+, \quad \ln O(D|\overline{R}) = \ln O(D) + W^-$$

These equations above are related to Bayes theorem: $O(D)$ can be thought of as the naive probability (i.e. raw probability without any conditioning), $W^{+,-}$ the likelihood or updating information and $O(D|R)$ the posterior probability.

Rearranging terms, we obtain the definition of *weights of evidence*:

$$W^+ = \ln O(D|R) - \ln O(D) = \ln\left[\frac{O(D|R)}{O(D)}\right] = \ln\left[\frac{T_{11}/T_{01}}{T_{1.}/T_{0.}}\right] = \ln\left[\frac{T_{11}T_{0.}}{T_{01}T_{1.}}\right]$$

and

$$W^- = \ln O(D|\overline{R}) - \ln O(D) = \ln\left[\frac{O(D|\overline{R})}{O(D)}\right] = \ln\left[\frac{T_{10}/T_{00}}{T_{1.}/T_{0.}}\right] = \ln\left[\frac{T_{10}T_{0.}}{T_{00}T_{1.}}\right]$$

2 Actual data from rasters

	R	$\overline{R}$	T
D	$247,116,300$	$161,730,900$	$408,850,200$
$\overline{D}$	$208,665,000$	$581,883,300$	$790,548,300$
T	$455,784,300$	$743,614,200$	$1,199,398,500$

The naive probability of deforestation, which can be calculated by summing D over $(R, \bar{R})$ yielding the marginal $P(D)$ is:

$$P(D) = \frac{408,850,200}{1,199,398,500} = 0.341$$

The deforestation odds $O(D)$ are:

$$O(D) = \frac{P(D)}{1 - P(D)} = 0.517$$

The posterior odds $O(D|R)$ are:

$$O(D|R) = \frac{247,119,300}{208,665,000} = 1.184$$

The posterior logit is $\ln(1.184) = 0.169$ and the 'naive' logit is $\ln(0.517) = -0.659$. Thus, $W^+ = 0.829$

What are the odds of deforestation outside the 1 km buffer?

$$O(D|\overline{R}) = \frac{161,730,900}{581,883,300} = 0.278$$

The posterior logit is $\ln(0.278) = -1.280$; the 'naive' logit is the same as before; and $W^- = -0.621$. The contrast $C_W \equiv W^+ - W^- = 1.449$.

Now, with those values in hand, we can 'reverse engineer' those logits into odds and into posterior probabilities.

	Road buffer	Posterior logit	Posterior odds	Posterior probab.
Case 1	Inside	0.169	1.184	0.542
Case 2	Outside	-1.280	0.278	0.217

Posterior probability follows from the fact that odds θ are $\theta = \frac{p}{1-p}$. Solving for p, we get $p = \frac{\theta}{1+\theta}$.

Compare these posterior probabilities with the naive probability. Notice that by including information about whether deforestation is inside or outside the 1 km buffer from roads, we improved our probability estimates.

Last Modified: November 30, 2018