Spatial Forecasting of Regional States

Lee De Cola and Stephen C. Guptill
U.S. Geological Survey
Reston, VA 20192 USA
ldecola@usgs.gov, sguptill@usgs.gov

Many geographic data are counts of events (photons in pixels, trees in quadrats, infection reports in counties), whose space-time density in regions is presumed to reflect the varying intensity of fields (radiation, vegetation, disease risk) within a universe E. Let $A \subset E$ be a space-time partitioning into N regions and T time-slices, so that $A = \cup A_{ij}$ where $\{(i\ ,j)\} = \{1, ..., N\} \times \{1, ..., T\}$. We let $k = 1,...,M$ measurements be made then $\boldsymbol{Z}(A) = Z_k(i, j)$ is an $N{\times}T{\times}M$ data matrix of counts recorded at each A_{ij}. Assume that these data represent a vector-valued field $\boldsymbol{\Phi} = \{\Phi_k : k = 1, ..., M\}$ on a d+1-dimensional space-time continuum $E = \{(x, t)\colon x \in \mathrm{Y}^d, t \in (0, \infty)\}$ and let $\boldsymbol{Z}(x, t)$ be a model of $\boldsymbol{\Phi}(x, t)$.

Prediction uses data $\boldsymbol{Z}(i, j)$ to parameterize the model $\boldsymbol{Z}(A)$ in order to understand the structure and behavior of the field $\boldsymbol{\Phi}(E)$ by assembling a system whose outputs 1) should reflect what we theorize to be the behavior of $\boldsymbol{\Phi}$, 2) yield small residuals $|(\boldsymbol{Z}(i, j) - \boldsymbol{Z}(x, t)|$ when $|(i, j) - (x, t)|$ is small, and 3) be visualizable in the sense that the output of $\boldsymbol{Z}(A)$ can be represented in tables, graphics, maps, animations—as well as uncertainty descriptions—useful to people who want to know more about $\boldsymbol{\Phi}$. Forecasting is a type of modeling that provides temporal predictions $\boldsymbol{Z}$ about $\boldsymbol{\Phi}$: $\boldsymbol{Z}(i, t) = F(\boldsymbol{Z}(i, j))$, where $F()$ is a model system encompassing not only future values at time $t > T$ but also information about the quality of those values. Interpolation alternatively estimates values over space: $\boldsymbol{Z}(x, t) = F(\boldsymbol{Z}(i, t))$, particularly at locations where $x \neq i$. Finally spatial forecasting $\boldsymbol{Z}(x, t) = F(\boldsymbol{Z}(i, j))$ is intended to produce models $\boldsymbol{Z}(A)$ that represent not only what are likely to be the values of $\boldsymbol{\Phi}(A)$ but also our confidence in those forecasts.

We narrow the above framework by assuming that the spatial dimension $d = 2$ and that Z is a binary-valued 2-dimensional vector ($M = 2$ and $Z \in [0, 1]$), e.g. representing whether some threshold level of counts or density has been exceeded. If we further assume that T = 2, then $\boldsymbol{Z}$ is simply two binary maps describing whether the threshold is exceeded within a region at a given time. A key forecasting issue is the association of $Z(i, t+1)$ with $Z(i, t)$:

		$t = 2$	
		1	0
$t = 1$	1	N_{11}	N_{12}
	0	N_{21}	N_{22}

We characterize the above contingency table, where $N_{11}+N_{12}+N_{21}+N_{22} = N$, with four indices with varying symmetry, scaling, and informational qualities. First is $p(\chi^2 \geq X^2)$, the exceedance probability of the chi-square statistic, which quantifies our surprise at the deviation from expectations under independence. The index is symmetric with respect not only to time ($p_{\chi 2}$ is unchanged when the numbers in the table are transposed and therefore temporal precedence is ignored) but also to positivity (it is unaffected if rows *and* columns are switched). This scale- and area-modification-dependent probability reflects the degree of temporal association.

A second index is r, the correlation coefficient of the above matrix viewed as a regression of $Z(i, 2)$ on $Z(i, 1)$. This index is also symmetric with respect to time, but not with respect to positivity, and it is not scale-dependent (in the sense that multiplying the cells of the table by a constant leaves r unchanged), and because $-1 \le r \le 1$ it is a useful summary measure of whether there appears to be an association (positive, or negative) between regional states at two time instants.

The next two indices, although quite general, have epidemiological interpretations. The third index is simply N_{11} the “true positives” in the table, and this is a useful summary statistic not only because $0 \le N_{11}/N \le 1$ but also because it reflects what in biomedicine is called the “sensitivity” of a test (just as N_{22} reflects “specificity”). A fourth index is $RR = (N_{11}/(N_{11}+N_{12})) \;/\; (N_{21}/(N_{21}+N_{22}))$, which describes the extent to which regional states at $t = 1$ forecast states at $t = 2$, suggests what is sometimes called “relative risk” by comparing the proportion of true positives among predicted positives with the proportion of false negatives among predicted negatives. Although not sensitive to scale, because $RR \in [0, \infty)$ it can be made more tractable by various transformations.

It is obvious that even so simple a situation as a single binary-valued phenomenon measured at 2 time instants ($N \times T \times M = N \times 2 \times 2$) provides a rich context for research. Compare West Nile virus reports for 3141 counties × 2 weeks × 2 “taxa” (birds and humans).

		t = week 33	
		HUMANS	
	U.S. STATES	POSITIVE	NEGATIVE
t = week 29	POSITIVE	204	363
BIRDS	NEGATIVE	386	2188

Consider each of the indices for the above table, aggregated from a (3141-county × 156-week × 3-taxon ≈) 1.5-million cell data matrix. First we can see that the association is almost certainly not random and the fact that $\chi^2 = 132.7$ ($p_{\chi 2} \sim 0$) confirms this; and because $r = 0.207$ the relationship between the states is positive and moderately strong. Finally $RR = 2.40$, i.e. if a region reported infected birds in 2001 then it had a 140% greater chance of reporting infected humans in 2002 than states where no West Nile Positive birds were reported (note that this language assumes away a number of epidemiological caveats).

This research, which is driven by public health concerns, raises several fundamental GIScience questions, including: What shall be the appropriate universe of analysis? How can we visualize and provide to decisionmakers the plethora of results when *A* grows to encompass linked, and often quite sparse and messy databases? How shall we account for, understand, and abstract from temporal and spatial autocorrelation as well as uncertainty? What do we gain and lose from mingling the language of regions (areas that can be modified and proliferated), objects (organisms, infections), and fields (risk surfaces in infinite spaces)? How can we manage scale and geometry in order to reliably interpolate, extrapolate, generalize and otherwise transform results? Finally, because this geography speaks of how patterns here and now in one phenomenon influence patterns there and then in another, can it help move us toward a mature understanding of future threats in an increasingly risk-preoccupied world?