

Statistical Assessment of Departure from Historical Conditions¹

Brian Steele and Swarna Reddy
Dept. of Mathematical Sciences
University of Montana

Introduction

- The long version of this talk is contained in a pdf file at www.math.umt.edu/steele/RowProjTalk.pdf
- Objective: develop a statistical method for assessing departure of current conditions from historical (reference) conditions.
- Specifically, use the output from LANDSUM (Keane et al. 2002²) and data on current conditions to measure departure and compute the observed significance levels (p-values)

Set-up

- The base map is a lattice of 30 m² pixels
- Every pixel belongs to a single potential vegetation type (PVT). Succession class varies over time in response to succession and fire events according to PVT-specific multiple pathway models
- The reporting unit is a 1 km² stratum (≈1100 pixels/stratum) ⇒ each stratum is separately examined for departure
- Simulations are sampled at 20 to 50 year intervals over 4,000 to 10,000 simulation years. $n = 200$ observations on successional class distribution are used to characterize reference conditions

Departure

- Departure from historical conditions means that the current distribution of successional classes, across the strata, is atypical of reference conditions.
- Measuring departure has been the difficulty

Significance

- Computing the observed significance level is straightforward
- However, there are limitations on the type of inferences that can be drawn because of lack of independence both temporally and spatially

¹JVA # 03-JV-11222048-151, LANDFIRE STATPAK

²Keane, R.E., Parsons, R.A., Hessburg, P.F. 2002. Estimating historical range and variation of landscape patch dynamics: limitations of the simulation approach. *Ecological Modelling*, **151**, 29-49.

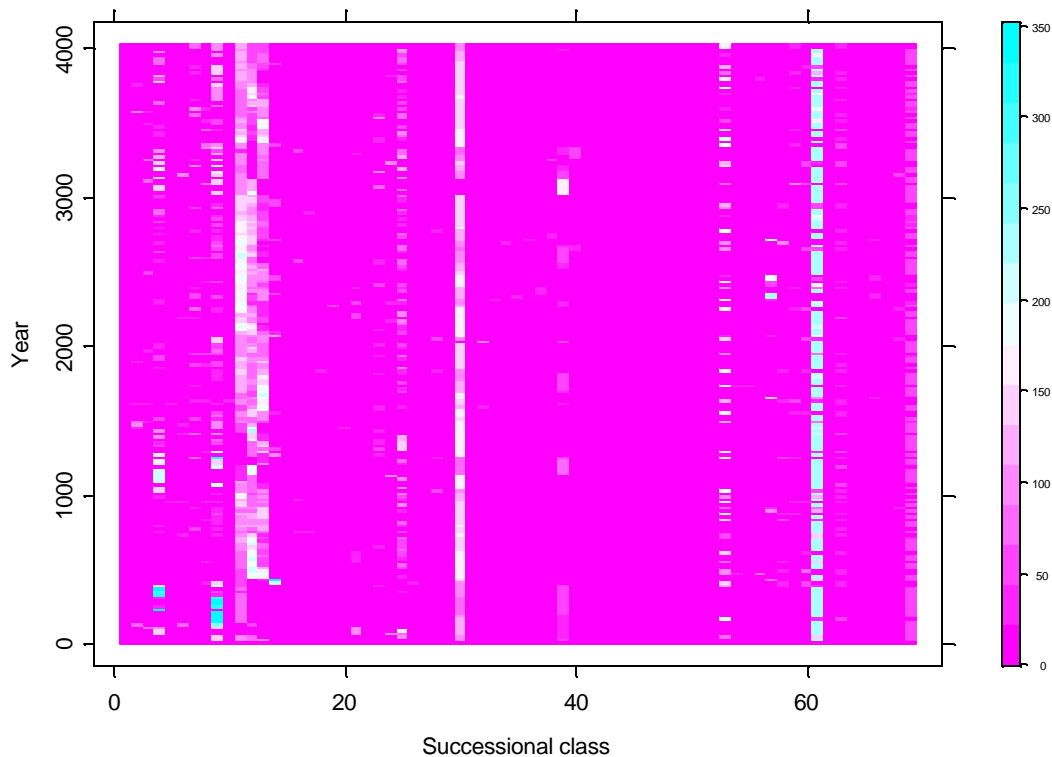
A Look at the Data

- The data for a particular stratum are analyzed as an $n \times p$ table denoted by \mathbf{X} where

$$n = \# \text{ years}$$

$$p = \# \text{ PVT/Successional class combinations}$$

- $50 \leq p \leq 250$
- $200 \leq n \leq 1000$
- A level plot shows the distribution of pixels in each possible PVT/successional class (horizontal axis) versus observation year (vertical axis)



- From a statistical perspective, the data consist of n observations on a multivariate stochastic process
- The $n \times p$ data matrix \mathbf{X} is not full rank. This has posed some computational difficulties

Measuring Departure

- The problem of comparing a single observation to a set of observations, and measuring departure is similar to the problem of *multivariate outlier detection*
- Let \underline{x}_0 denote an observation on current conditions for some strata.

- \underline{x}_0 is a p -vector consisting of the number of pixels in each of the successional classes occurring on the stratum
- Mahalanobis distance is a commonly-used measure of departure that compares \underline{x}_0 to a sample mean:

$$d_M = (\underline{x}_0 - \bar{\underline{x}})^T \mathbf{D}^{-1} (\underline{x}_0 - \bar{\underline{x}})$$

where $\bar{\underline{x}}$ is the p -vector of reference means computed from \mathbf{X} , \underline{x}^T denotes the transpose of \underline{x} , and \mathbf{D} is the $p \times p$ sample variance matrix computed from \mathbf{X}

- Using Mahalanobis distance will reduce the impact of successional classes with large variances
- The treatment of successional classes according to variance is a poor property in this situation because high variance successional classes also are the most common classes
- Because \mathbf{X} is not full rank, \mathbf{D}^{-1} does not exist, and we use the Moore-Penrose (generalized) inverse in place of \mathbf{D}^{-1}

The Row Projection Measure of Departure

- We propose an original measure of departure: the row projection measure
- The row projection measure finds the best possible linear approximation of \underline{x}_0 using the observations in \mathbf{X} , and measures the approximation error
- If \underline{x}_0 is similar to the observations in \mathbf{X} , then the error will be small; if \underline{x}_0 is not similar to the observations in \mathbf{X} , then the error will be large
- The linear approximation of \underline{x}_0 is found by projecting \underline{x}_0 onto the *row* space of \mathbf{X}
- The approximation error (and departure) is measured by

$$\varepsilon_0^2 = 1 - \underline{x}_0^T \mathbf{U}_r \mathbf{U}_r^T \underline{x}_0 \quad (1)$$

where \mathbf{U}_r are the left-eigenvalues of the truncated singular value decomposition of \mathbf{X}

- $\mathbf{U}_r \mathbf{U}_r^T$ is the projection matrix for the row space of \mathbf{X}
- It can be proven that

$$\varepsilon_0^2 = 1 - R^2, \quad ,$$

where R^2 is the *multiple correlation coefficient* obtained from the linear regression of \underline{x}_0 on the rows of \mathbf{X}

- Thus, $100\varepsilon_0^2$ is the percentage of the variation in \underline{x}_0 that cannot be explained by \mathbf{X}
- An alternative method projects \underline{x}_0 onto the p -vector of sample means $\bar{\underline{x}}$ computed from \mathbf{X} and measures the lack-of-fit associated with approximating \underline{x}_0 by $\bar{\underline{x}}$. This measure is

$$\begin{aligned}\eta_0^2 &= 1 - (\underline{x}_0^T \bar{x})^2 / (\bar{x}^T \bar{x}) \\ &= 1 - r^2\end{aligned}$$

where r is the sample correlation coefficient between \underline{x}_0 and \bar{x}

- The measure η_0^2 generally is not as sensitive to departures as is ε_0^2

Significance Testing

- Let \mathcal{P} denote the stochastic process generating the rows of \mathbf{X}
- Let

$$T = T(\underline{x}_0, \mathbf{X})$$

denote the statistic generating the observed departure ε_0^2 . The distribution of T is unknown

- We use a distribution-free approach to test

$$H_0 : \underline{x}_0 \in \mathcal{P} \text{ versus } H_1 : \underline{x}_0 \notin \mathcal{P}$$

- The distribution of T under H_0 is estimated by resampling as follows
- Assume that H_0 is true. Then

$$\mathbf{X}_0 = \begin{pmatrix} \underline{x}_0^T \\ \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix}_{(n+1) \times p}$$

is a random sample from \mathcal{P}

- The resampling estimate of the distribution of T is computed by randomly drawing a row from \mathbf{X}_0 and computing the departure of the sampled row from the remaining rows
- This process is repeated until a sufficiently large set of observations on departure are generated, say $\hat{F} = \{\varepsilon_1^2, \dots, \varepsilon_m^2\}$
- The observed significance level is the proportion of simulated values that are larger than the observed value ε_0^2 :

$$\text{p-value} = \hat{P}[T \geq \varepsilon_0^2 \mid H_0] = \frac{\#\{\varepsilon_k^2 \in \hat{F} \mid \varepsilon_k^2 \geq \varepsilon_0^2\}}{\#\{\varepsilon_k^2 \in \hat{F}\}}$$

- However, there are only $n + 1$ possible ways to draw one row from $n + 1$. Consequently, we construct all possible partitions, and compute

$$\hat{F} = \{\varepsilon_0^2, \varepsilon_1^2, \dots, \varepsilon_n^2\}$$

A Comparison of Departure Measures

- A simulation study was conducted to compare the relative sensitivity of the departure measures discussed above
- The data are presented in sets of 256 spatially contiguous strata. To obtain outliers to contaminate the i th stratum data set, we draw a sample of r observations from other strata and combine with the i th stratum data set
- If the contamination rate is $p\%$, then we identify the largest $p\%$ of the departures, and flag those observations as outliers.
- Among that set of identified outliers, we determine the fraction of observations that are true outliers. This is the *outlier detection rate*. The percentage that are not outliers is called the *false positive rate*
- The results shown below for LH50K2 are representative of the four examined data sets

Figure 2. Outlier detection rate plotted against contamination fraction for four measures of departure. Data set is LH50K2. Plotted values are averages over 256 data sets (strata)

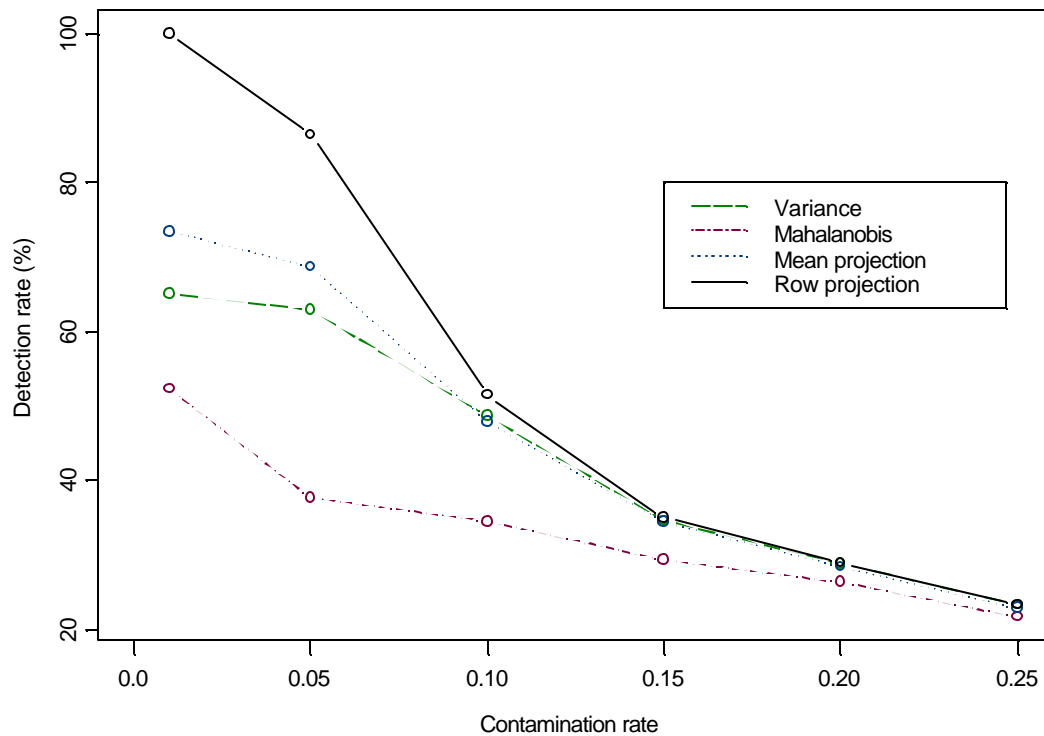
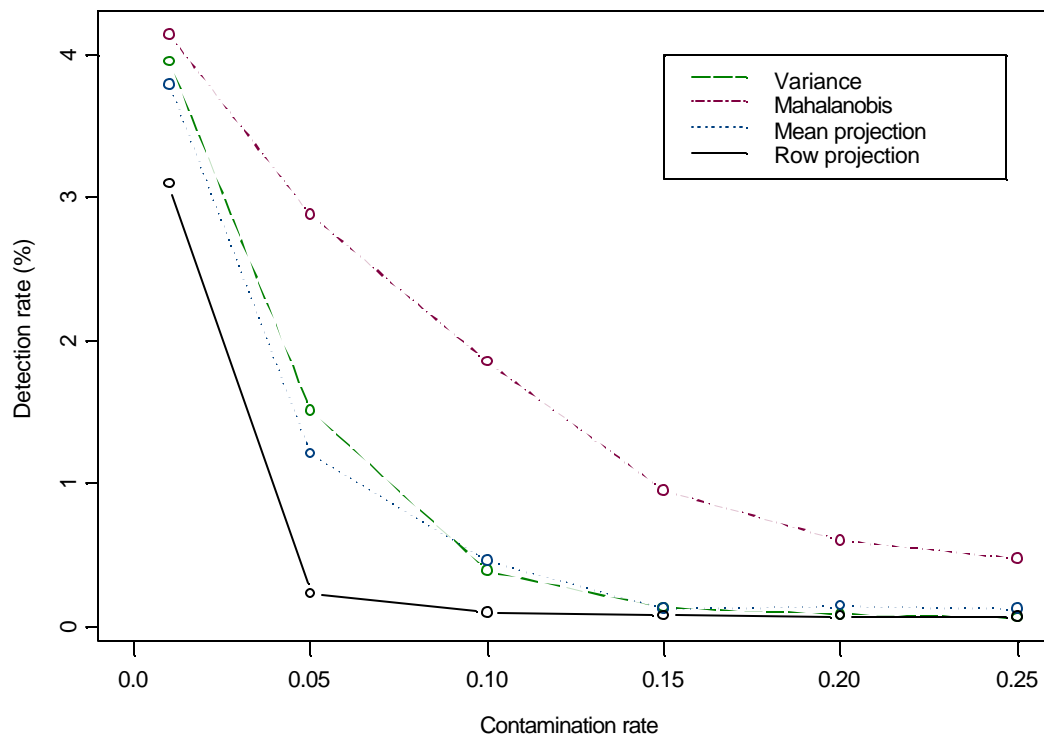


Figure 3. False positive rate plotted against contamination fraction for four measures of departure. Data set is LH50K2. Plotted values are averages over 256 data sets (strata).



Maps

- A map of the p-values shows that for most the area, there is substantial *evidence* of departure from reference conditions
- A map of the FRCC shows most of the area in the most extreme of the three categories

Conclusions

- The row projection method provides a sensitive test of departure from reference conditions
- A stratum is likely to be identified as significantly different from reference conditions if the simulated data are not representative of reference conditions

Caveats

- The maps show the p-value - a measure of *evidence* of departure, not a measure of departure
- The maps are based on comparison of simulation data to current conditions. We have assumed that the simulation is accurate - if not, then the maps are not reliable
- Small p-values do not necessarily imply $FRCC = 3$. For example, if a stratum is dominated by Spruce/Fir historically, and current conditions are dominated by Aspen-Birch, then the FRCC will be Class 3

