

Statistical Assessment of Departure from Historical Conditions¹

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

Introduction

- From URL: <http://www.landfire.gov/>: *Over ninety years of fire exclusion, domestic livestock grazing, logging and widespread exotic species invasions have altered fire regimes, fuel loadings, and vegetation composition and structure. As the result, the number, size, and intensity of wildfires have significantly changed from the historic conditions, sometimes with catastrophic consequences.*
- *In response to these severe conditions, the federal government has developed a National Fire Plan, and the Department of Agriculture Forest Service and the Department of the Interior are jointly conducting a cohesive strategy to implement the Plan.*
- *The LANDFIRE project is a multi-agency, inter-disciplinary research and development activity designed to develop a consistent and accurate methodology capable of producing geospatial data of vegetation conditions, fire fuels, risks, and ecosystem status at the national, regional, and local scales for implementation of the National Fire Plan.*
- Objective: develop a statistical method for assessing departure from historical conditions.
- LANDSUM (Keane et al. 2002²) is used to simulate reference conditions (historical data) across a landscape.
- There are two map units: 30 m² pixels, and strata (1 km²)
- Every 30 × 30 m 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
- Simulations are sampled at 20 to 50 year intervals over 4000 to 10000 simulation years
- Current conditions (successional class) are to be compared to reference conditions for each strata
- Departure of current from historical conditions is assessed for each strata

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.

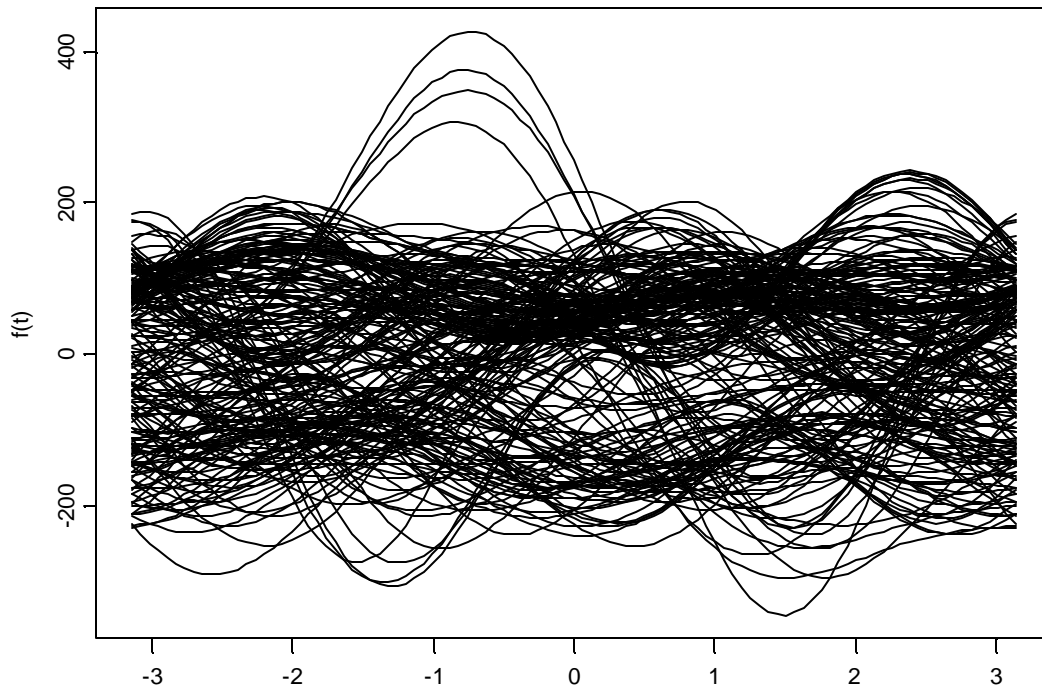
A Look at the Data

- The data for a particular stratum are analyzed as an $n \times c$ table.

$n = \# \text{ years}$

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

- $50 \leq p \leq 250$
- $200 \leq n \leq 1000$
- A level plot shows the distribution of pixels in each possible PVT/successional class (horizontal axis) versus observation year (vertical axis) (See Figures)
- The data are inherently multidimensional. An Andrews plot (below) shows the patterns of variation across years. Each curve corresponds to an observation year (or row). There are 4 or 5 identifiable patterns.



- From a statistical perspective, the data consist of n observations on a multivariate stochastic process
- From a computational perspective, the $n \times p$ data matrix \mathbf{X} is difficult to work with because the columns are not linearly independent. The rank of \mathbf{X} is substantially less than p

Objectives

- The primary objective is to develop a method of detecting observations that are unusual with respect to the stochastic process that generated \mathbf{X}

- The method will then be used to test whether current conditions have departed from historical conditions
- The problem is most similar to that of *detecting multivariate outliers*
- Most *multivariate outlier detection* methods are based on reducing the dimensionality of the data by extracting several axes of variation and measuring the extent to which an observation is an outlier with respect to these axes
- Dimension reduction is carried out via the spectral decomposition of the $p \times p$ dispersion matrix (or sample variance) D given by

$$D = V \Lambda V^T$$

$$= \sum_{i=1}^p \lambda_i \underline{v}_i \underline{v}_i^T$$

where $V = (\underline{v}_1 \ \underline{v}_2 \ \cdots \ \underline{v}_p)$ is the eigenvector matrix and $\Lambda = \text{diag}(\lambda_1 \ \lambda_2 \ \cdots \ \lambda_p)$ is a diagonal matrix comprised of the eigenvalues of D

- Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$
- The principal components

$$\begin{aligned} y_1 &= X \underline{v}_1 \\ &\vdots \\ y_p &= X \underline{v}_p \end{aligned}$$

are the projections of the data onto each of the axes defined by the eigenvectors

- Let \underline{x}_0 denote an observation suspected of being an outlier
- One approach measures the length of the projection of $\underline{x}_0 - \bar{\underline{x}}$ onto either the first r axes or the last s axes. For example, if $V_r = (v_1 \ v_2 \ \cdots \ v_r)$ and $\Lambda_r = \text{diag}(\lambda_1 \ \lambda_2 \ \cdots \ \lambda_r)$, then the projection onto these axes is

$$V_r(\underline{x}_0 - \bar{\underline{x}}).$$

- The squared length of the projection is

$$d_0^1 = (\underline{x}_0 - \bar{\underline{x}})^T V_r^T V_r (\underline{x}_0 - \bar{\underline{x}})$$

- If $\bar{\underline{x}} \sim (\bar{\underline{x}}, D)$, then the variance of the projection is

$$\text{Var}[V_r(\underline{x}_0 - \bar{\underline{x}})] = V_r^T D V_r = V_r^T V \Lambda V^T V_r = \Lambda_r$$

- If \underline{x}_0 inflates the *sample* variance with respect to one or more of the principal axes, then it is likely to be an outlier according to d_0^1 . That is, it must have an extreme value on at least one of these axes

- Another approach is to use the last s axes (associated with the smallest s eigenvalues). This approach will identify observations that are not identifiable by other methods (in principle)
- For the problem at hand, it is difficult to identify the last meaningful s axes because \mathbf{X} is not full rank. The smallest non-zero eigenvalues may or may not be associated with real axes of variation
- A more familiar outlier detection measure is Mahalanobis distance

$$d_0^3 = (\underline{x}_0 - \bar{\underline{x}})^T \mathbf{D}^{-1} (\underline{x}_0 - \bar{\underline{x}})$$

- Because \mathbf{X} is not full rank, \mathbf{D}^{-1} does not exist, and we use the Moore-Penrose (generalized) inverse

$$\mathbf{D}^- = \mathbf{V}_k \mathbf{\Lambda}_k^{-1} \mathbf{V}_k^T$$

where k is the rank of \mathbf{X} , in place of \mathbf{D}^{-1} . Then,

$$d_0^3 = (\underline{x}_0 - \bar{\underline{x}})^T \mathbf{V}_k \mathbf{\Lambda}_k^{-1} \mathbf{V}_k^T (\underline{x}_0 - \bar{\underline{x}})$$

- Mahalanobis distance gives each variable the same weight (in terms of variance) and reduces the combined effect of variables that are positively correlated
- The central idea behind these measures is to identify observations that are distant from the centroid defined by the data \mathbf{X} . The distance from \underline{x}_0 to the centroid is adjusted to reflect the dispersion along an axis in the direction from \underline{x}_0 to the centroid

The Row Projection Measure of Departure

- We propose an approach that decomposes \underline{x}_0 as a sum of orthogonal vectors \underline{u} and \underline{v} :

$$\underline{x}_0 = \underline{u} + \underline{v} \text{ where } \underline{u}^T \underline{v} = 0$$

- \underline{u} is the projection of \underline{x}_0 onto the *row* space of \mathbf{X} . Hence, \underline{u} is a linear combination of the rows of \mathbf{X}
- \underline{v} is the error, i.e., $\underline{v} = \underline{x}_0 - \underline{u}$
- The length of \underline{v} reflects the extent to which \underline{x}_0 is an outlier
- A projection matrix for the column space $\mathcal{C}(\mathbf{X})$ spanned by the columns of \mathbf{X} is idempotent, symmetric and satisfies $\mathbf{P}\mathbf{X} = \mathbf{X}$
- The projection matrix for $\mathcal{C}(\mathbf{X})$ can be expressed as

$$\mathbf{P}_{n \times n} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

- The projection of an observation \underline{x}_0 onto \mathbf{X} given by $\hat{\underline{x}}_0 = \mathbf{P}\underline{x}_0$ is an approximation that is closest in terms of the squared error

$$\varepsilon_0^2 = \sum_{i=1}^p (x_{i,0} - \hat{x}_{i,0})^2 = \underline{x}_0^T (\mathbf{I}_n - \mathbf{P}) \underline{x}_0$$

• In other words, there is no other approximation derived from \mathbf{X} that can be closer. The least squares prediction of \underline{y} when \mathbf{X} is full rank is an example:

$$\begin{aligned} \underline{\hat{y}} &= \mathbf{P} \underline{y} \\ &= \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underline{y} \\ &= \mathbf{X} \underline{\hat{\beta}} \end{aligned}$$

where $\underline{\hat{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underline{y}$ is the least squares estimator for the model $\underline{Y} = \mathbf{X} \underline{\beta} + \underline{\varepsilon}$

• Because \mathbf{P} is idempotent, the squared length of \underline{x}_0 can be written as

$$\begin{aligned} \underline{x}_0^T \underline{x}_0 &= \underline{x}_0^T (\mathbf{P} + \mathbf{I}_n - \mathbf{P}) \underline{x}_0 \\ &= \underline{x}_0^T \mathbf{P} \underline{x}_0 + \underline{x}_0^T (\mathbf{I}_n - \mathbf{P}) \underline{x}_0 \\ &= \underline{x}_0^T \mathbf{P} \underline{x}_0 + \varepsilon_0^2 \end{aligned}$$

• Our measure of departure 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} , or equivalently, the column space $\mathcal{C}(\mathbf{X}^T)$.

• If \mathbf{X}^T is full rank, then the projection matrix can be computed as

$$\mathbf{P}_{p \times p} = \mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} \tag{1}$$

• However, $n > p \Rightarrow (\mathbf{X} \mathbf{X}^T)^{-1}$ does not exist. However, the projection matrix does exist. To derive the projection matrix, consider the singular value decomposition of \mathbf{X}^T given by

$$\mathbf{X}_{p \times n}^T = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T$$

where $\mathbf{U}_{p \times p}$ is comprised of the eigenvectors of $\mathbf{X}^T \mathbf{X}$ and $\mathbf{V}_{n \times n}$ is comprised of the eigenvectors of $\mathbf{X} \mathbf{X}^T$. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ denote the n eigenvalues of $\mathbf{X} \mathbf{X}^T$ (the first p of which are also the eigenvalues of $\mathbf{X}^T \mathbf{X}$); then $\mathbf{\Lambda}$ denotes the diagonal matrix with diagonal elements $\lambda_k^{1/2}$, $k = 1, \dots, n$.

- Suppose that the rank of \mathbf{X} is $r \leq p$. Then, there are $n - r$ singular values (eigenvalues equal to 0), and \mathbf{X}^T can be expressed in terms of the truncated singular value decomposition

$$\begin{aligned}\mathbf{X}^T &= \mathbf{U} \begin{pmatrix} \mathbf{\Lambda}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V}^T \\ &= (\mathbf{U}_r \quad \mathbf{U}_{n-r}) \begin{pmatrix} \mathbf{\Lambda}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-r} \end{pmatrix} \begin{pmatrix} \mathbf{V}_r^T \\ \mathbf{V}_{n-r}^T \end{pmatrix} \\ &= \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^T\end{aligned}$$

- The Moore-Penrose generalized inverse of $\mathbf{X}\mathbf{X}^T$ is

$$\begin{aligned}(\mathbf{X}\mathbf{X}^T)^- &= \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^T \\ &= \mathbf{V}_r\mathbf{\Lambda}_r^{-2}\mathbf{V}_r^T\end{aligned}$$

where $\mathbf{\Lambda}^{-2}$ is defined to be the diagonal matrix with diagonal elements λ_i^{-1} if $\lambda_i > 0$ and 0 if $\lambda_i = 0, i = 1, \dots, n$

- Substituting $\mathbf{U}_r\mathbf{\Lambda}_r\mathbf{V}_r^T$ for \mathbf{X}^T and $\mathbf{V}_r\mathbf{\Lambda}_r^{-2}\mathbf{V}_r^T$ for $(\mathbf{X}\mathbf{X}^T)^{-1}$ in formula (1) leads to

$$\begin{aligned}\mathbf{P} &= \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{V}_r^T\mathbf{V}_r\mathbf{\Lambda}_r^{-2}\mathbf{V}_r^T\mathbf{V}_r\mathbf{\Lambda}_r\mathbf{U}_r^T \\ &= \mathbf{U}_r\mathbf{U}_r^T\end{aligned}$$

- To verify that $\mathbf{U}_r\mathbf{U}_r^T$ is the projection matrix onto the row space of \mathbf{X}^T , note that $\mathbf{U}_r\mathbf{U}_r^T$ is idempotent and symmetric, and

$$\begin{aligned}\mathbf{U}_r\mathbf{U}_r^T\mathbf{X}^T &= \mathbf{U}_r\mathbf{U}_r^T\mathbf{U}_r\mathbf{\Lambda}_r\mathbf{V}_r^T \\ &= \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{V}_r^T \\ &= \mathbf{X}^T.\end{aligned}$$

- $\mathbf{I}_p - \mathbf{U}_r\mathbf{U}_r^T$ is the projection matrix onto the orthogonal complement $\mathcal{C}(\mathbf{X}^T)^\perp$ of the row space of \mathbf{X}^T . The squared error incurred by approximating \underline{x}_0 by the projection

$\mathbf{P}\underline{x}_0$ is

$$\begin{aligned}\varepsilon_0^2 &= \underline{x}_0^T(\mathbf{I}_n - \mathbf{P})\underline{x}_0 \\ &= \underline{x}_0^T\underline{x}_0 - \underline{x}_0^T\mathbf{P}\underline{x}_0\end{aligned}$$

- It is convenient to normalize \underline{x}_0 to have unit length. Then, $\underline{x}_0^T\underline{x}_0 = 1$, and the departure of \underline{x}_0 from the row space of \mathbf{X}^T is measured by

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

- Notes

1. $0 \leq \varepsilon_0^2 \leq 1$

2. It is not necessary to compute the singular value decomposition of \mathbf{X} because \mathbf{U}_r is constructed from the eigenvectors of the $p \times p$ matrix $\mathbf{X}^T \mathbf{X}$. Hence, the computational demand is minor

- A simple alternative method, not requiring the spectral decomposition of $\mathbf{X}^T \mathbf{X}$, compares \underline{x}_0 to $\bar{\underline{x}}$ (or more generally, any centroid of \mathbf{X})

- For example, project \underline{x}_0 onto $\mathcal{C}(\bar{\underline{x}})$ and measure the lack-of-fit. The projection matrix for $\mathcal{C}(\bar{\underline{x}})$ is $\bar{\underline{x}}(\bar{\underline{x}}^T \bar{\underline{x}})^{-1} \bar{\underline{x}}^T$ and the departure of \underline{x}_0 from $\mathcal{C}(\bar{\underline{x}})$ is

$$\eta_0^2 = 1 - (\underline{x}_0^T \bar{\underline{x}})^2 / (\bar{\underline{x}}^T \bar{\underline{x}})$$

- In this situation, we are measuring departure with respect to a single axis specified by the mean vector $\bar{\underline{x}}$

- Notice that

$$(\underline{x}_0^T \bar{\underline{x}})^2 / (\bar{\underline{x}}^T \bar{\underline{x}}) = \cos^2(\underline{x}_0, \bar{\underline{x}}) = r^2(\underline{x}_0, \bar{\underline{x}})$$

(squared Pearson's correlation)

- Theoretically, a robust measure of departure can be constructed by replacing $r^2(\underline{x}_0, \bar{\underline{x}})$ with a robust correlation coefficient

Significance Testing

- Let $t(\mathbf{X}_0)$ denote the statistic generating the observed departure ε_0^2 . The distribution of $t(\mathbf{X}_0)$ is not known, so a distribution-free test is used to test

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

where \mathcal{P} is the stochastic process generating the rows of \mathbf{X}

- We adopt a resampling approach to estimating the distribution of $t(\mathbf{X}_0)$ under H_0 (or any of the distance measures discussed above). Under H_0 , $\mathbf{X}_0 = \{\underline{x}_1, \dots, \underline{x}_n, \underline{x}_0\}$ is a random sample from a single population, or stochastic process.

- To estimate the observed significance level, $P[t(\mathbf{X}_0) \geq \varepsilon^2 \mid H_0]$, we randomly partition \mathbf{X}_0 as $\{\underline{x}_k\}$ and $\mathbf{X}_{-k} = \{\underline{x}_1, \dots, \underline{x}_{k-1}, \underline{x}_{k+1}, \dots, \underline{x}_n, \underline{x}_0\}$ and compute the departure ε_k^2 of \underline{x}_k from the row space of \mathbf{X}_{-k} .

- 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 value is then the proportion of departures that are larger than ε_0^2 ; i.e.,

$$\widehat{P}[t(\mathbf{X}_0) \geq \varepsilon^2 \mid H_0] = \frac{\#\{\varepsilon_k^2 \in \widehat{F} \mid \varepsilon_k^2 \geq \varepsilon_0^2\}}{\#\{\varepsilon_k^2 \in \widehat{F}\}}$$

- However, there are only $n + 1$ ways to split \mathbf{X}_0 so that one set is singleton. Consequently, we construct all possible partitions.

Swamping

- An important aspect of outlier detection is swamping. Swamping refers to a situation in which the proportion of outliers in a data set is so large as to substantially reduce the sensitivity of a detection method to outliers
- All of the research on outlier detection methods in the past 10 years has focussed on constructing methods that are robust against swamping. The computational complexity of these methods is substantial
- While swamping is not a concern in this study, it is an interesting aspect of test sensitivity

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 identify those observations as outliers.
- Among that set of 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*

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)

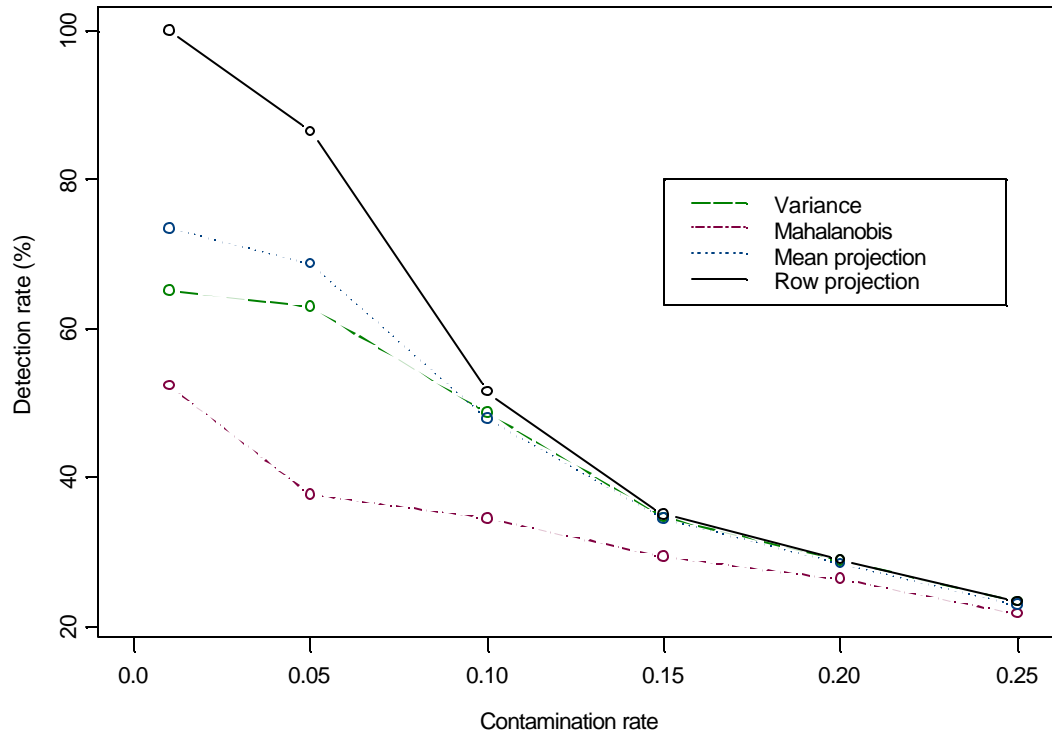


Figure 3. False positive rate plotted against contamination fraction for four measures of departure. Data set is LH50K2. Plotted values are averages over 256 data sets (strata).

