

Surveying Data for Patchy Structure

Ronald K. Pearson*

Abstract

The term “data surveying” refers to the preliminary examination of a dataset to assess its overall character, and this process typically involves simple descriptive statistics to characterize the available variables, along with detection of data anomalies (e.g., outliers or incomplete records) and possibly other “interesting” or “unusual” features that may be worthy of careful scrutiny. In the survey sampling literature, an important distinction is made between responses that are *missing at random*, the simplest form of *ignorable missing data*, and *non-random* alternatives that can lead to *non-ignorable missing data*. The distinction is practically important because non-ignorable missing data can cause severe biases in analytical results, while ignorable missing data typically causes an undesirable but less serious increase in the variability of these results. Analogous distinctions can also be usefully made for other types of data anomalies (e.g., i.i.d. vs. correlated outliers) or other unusual data subsets of potential interest. In particular, the observation of systematic behavior with respect to time, position, or other ordered index sequences (e.g., primary key in a database) can often give insight into the nature or generation mechanism of these data subsets. Motivated by these observations, this paper considers the problem of detecting structure in distinguished subsets of data records, including missing data, outliers and other “interesting data records.” Depending on the nature of the dependences considered, this problem is closely related to a number of others, including the detection of “streaks” in athletic performance records, the quantification of association between variables, or binary classification.

1 Introduction

Pyle [18] describes data mining in terms of three components: data preparation, data surveying, and data modeling. The second of these steps—data surveying—is concerned with identifying and characterizing what is present in the dataset. Useful data surveying tools include simple descriptive statistics (e.g., how many variables constitute each data record? what kind are they (nominal, ordinal, or real)? what are the ranges and typical values for each?), along with somewhat more complex characterizations like entropy measures [18, Ch. 11]. In addition, it is also useful to characterize

both data anomalies (e.g., outliers and missing data) and other “interesting” or “unusual” data subsets that may be worthy of separate analysis. The problem of interest in this paper is the detection of significant *structure* in these subsets, which may lead to useful insights concerning their nature and origin.

As a particularly important example, a useful distinction is made in the survey sampling literature between data values that are *missing at random (MAR)* and those that are *systematically missing* [20]. The MAR model generally represents *ignorable missing data*, which may be regarded as a nuisance that causes the uncertainty of our analytical results to increase, effectively reducing our sample size. Conversely, *nonignorable* missing data patterns in which the probability of being included in the dataset depends on the missing data values themselves are generally more serious as they can cause large biases in our results. Further, the identification of nonignorable missing data can be the first step in discovering *why* these data values are missing, which can have important practical implications.

Although it is not as widely discussed, a somewhat analogous distinction is that between outliers that are randomly distributed throughout the dataset and *dependent outliers*, sometimes known as “patchy outliers.” In particular, Davies and Gather [6] express concern that, “in almost all simulations in the literature the outliers are taken to be iid random variables.” To illustrate that such working assumptions are not always appropriate, they discuss a highly contaminated weather balloon dataset in which the outliers do not conform to this assumption. This distinction is important because outlier sequences of the same magnitude and concentration but with different dependence structures can have very different influences on dynamic characterizations like spectrum analysis or linear system identification [16]. Again, detection of systematic patterns in outliers or other data anomalies can be useful in determining the mechanisms and sources responsible for these anomalies.

The analogy between dependent outliers and systematic missing data becomes clear if we adopt the *replacement model* for outliers [14]. There, the sequence $\{y_k\}$ of available data samples is modelled as:

$$(1.1) \quad y_k = (1 - z_k)x_k + z_k o_k,$$

*ProSanos Corporation, Harrisburg, PA.

where $\{x_k\}$ is the nominal (i.e., uncontaminated) data sequence of interest, $\{o_k\}$ is a sequence of outlying values, and $\{z_k\}$ is a binary selection sequence, assuming the value $z_k = 0$ whenever the nominal data value is observed, and $z_k = 1$ whenever the outlying data value is observed. The outlier analog of the missing at random data model then corresponds to assuming that $\{z_k\}$ is an iid binary sequence (i.e., a sequence of Bernoulli trials). Similarly, the outlier analog of systematic missing data corresponds to the case where the binary sequence $\{z_k\}$ either exhibits a significant dependence structure (e.g., patchy outliers) or depends on other contaminated variables. In particular, *common mode effects* (e.g., partial system failures) can be responsible for the presence of outliers in several different variables simultaneously, again in violation of the random occurrence model. The influence of these strongly correlated outlier sequences in different variables can profoundly influence the results of otherwise reasonable joint characterizations like cross-correlation analysis [15, Sec. 8.1.2].

Useful distinctions can also be made between random and systematic occurrence of other types of anomalous, interesting, or unusual data records \mathcal{R}_k within a dataset \mathcal{D} . Specific examples include *inliers*, defined as observations that lie within the distribution of nominal (i.e., non-anomalous) data values, but which are in error [21], *near-duplicate records* (e.g., web documents) [3], or subsets of data records that have been deemed “interesting” by some quantitative *interestingness measure* [12]. DesJardins [7] notes that isolated inliers may not be a problem and may be almost indistinguishable from nominal data values, but that moderate-sized sets of inliers can have more serious analytical consequences. (It is important to note that the term “inlier” is sometimes used as a synonym for “nominal data” [13], quite distinct from the meaning assumed here.) Finally, another case where data records of particular interest are not randomly distributed throughout a dataset is the case of alarms in telecommunication network data [22]. There, different alarm sequences are known to be correlated and to occur in intermittent bursts; one of the key practical challenges is in extracting cause information from the complicated patterns generated by these related alarm sequences.

The general problem considered in this paper is the following one. We are given a class \mathcal{K} of data records $\{\mathcal{R}_k\}$ that are of particular interest. This class could consist of missing or incomplete records, outliers, inliers, near-duplicate records, or records identified on the basis of some interestingness measure, either objective or subjective [12]. The essential requirement here is that we have available a classification scheme that partitions data records into those belonging to class \mathcal{K} and those

not belonging to this class. Given this partitioning, define the *status sequence* $\{z_k\}$ as:

$$(1.2) \quad z_k = \begin{cases} 1 & \text{if } \mathcal{R}_k \in \mathcal{K} \\ 0 & \text{if } \mathcal{R}_k \notin \mathcal{K}, \end{cases}$$

generalizing the binary sequence $\{z_k\}$ on which the replacement outlier model (1.1) is based. An obvious extension of this idea would be to consider multiple classes of interesting data records, but this paper considers only a single class \mathcal{K} . The key problem of interest here thus reduces to the characterization and interpretation of the binary sequence $\{z_k\}$. As one reviewer noted, it is important to emphasize that the utility of the results obtained from analyses like those described here depends strongly on the accuracy of the classification procedure that generates the status sequence $\{z_k\}$. To keep the length of this paper manageable, the problem of missclassification is minimized here by considering cases like missing data where accurate construction of the sequence $\{z_k\}$ is straightforward.

2 The problem of assessing patchiness

The primary question considered in this paper is whether the records belonging to class \mathcal{K} occur randomly or systematically through the dataset \mathcal{D} , based on their record index k . If the dataset consists of a sequence of real values indexed by time and if the class \mathcal{K} corresponds to local outliers in this sequence, the question considered here reduces to one of determining whether these outliers are isolated or patchy. More generally, even if the records are much more complex and the record index has no obvious interesting interpretation, the detection of patchiness in the status sequence $\{z_k\}$ can lead us to discover unexpected structure in these records, a point illustrated in Sec. 6.

Despite the simplicity of the concept—records from class \mathcal{K} are either grouped together into patches or they are not—the practical assessment of patchiness in binary data sequences is harder than it sounds. This point is illustrated in Fig. 1, which shows a portion of the status sequence $\{z_k\}$ discussed in Sec. 6.2. Briefly, this sequence identifies a subset of adverse event incident reports in the U.S. Food and Drug Administration’s Adverse Event Reporting System (AERS) database in which an outcome of “death” is listed. The visual appearance of this binary sequence is strongly suggestive of patchiness, but it is desirable to have a quantitative characterization that permits us to objectively assess patchiness and quantify its extent.

The assessment of patchiness in status sequences is essentially the same as that of detecting “streakiness” in sports performance statistics. As a specific example, Albert and Williamson consider the question, “Was

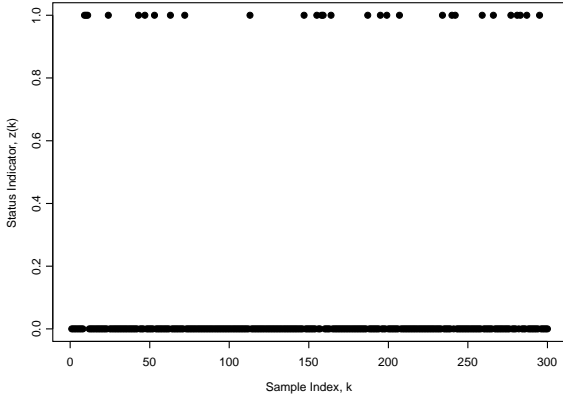


Figure 1: Binary status sequences derived from the AERS database example discussed in Sec. 6.2.

Javy Lopez a 'streaky hitter' when he played for the Atlanta Braves in 1998?" [1]. These authors discuss several different approaches to this problem, and the one they adopt is a simulation-based Bayesian strategy that incorporates any one of several different parametric streakiness models and draws inferences about the model parameters. In addition, this approach also requires a "streakiness statistic" and the authors consider six of these: two are based on moving averages, two are based on characterization of runs of successive 0's or 1's in individual at-bat results, one is based on a logistic model relating the probability of hitting in a given game to batting averages in previous games, and one is based on the standard deviation of batting averages across subgroups of games.

The approach to assessing patchiness adopted in this paper is based on the empirical patchiness measures described in Sec. 4, together with a random permutation strategy that provides a reference standard for tests of the patchiness hypothesis. To assess the performance of these measures, they are first applied to simulated sequences based on the patchy sequence model described in Sec. 3. One of these measures is then applied in Sec. 6 to the AERS database mentioned above.

3 A patchy sequence model

To assess the performance of patchiness characterizations like those described here, it is useful to have a simulation procedure for generating patchy sequences with well-defined, controllable characteristics. Here, the following patchy sequence model is used as the basis for such a procedure. The idea is to specify a distribution $\{p_w\}$ of patch widths w and use this distribution

to generate a binary sequence $\{z_k\}$ of length N having patches of successive 1's that are drawn from this distribution. Specifically, given $\{p_w\}$, the sequence $\{z_k\}$ is generated as follows. First, a sequence $\{w_k\}$ of N possible patch widths is generated having probabilities p_w for $w = 0, 1, \dots, w^*$ where $w^* \leq N$ is width of the widest patch considered. The binary sequence $\{z_k\}$ is then constructed according to the following procedure:

0. Initialize: set $k = 1$

1. Do while $k \leq N$:

- $w_k = 0 \Rightarrow z_k = 0$ and $k \rightarrow k + 1$,
- $w_k = 1 \Rightarrow z_k = 1$ and $k \rightarrow k + 1$,
- $w_k = w > 1 \Rightarrow z_k = z_{k+1} = \dots = z_{k+w} = 1$ and $k \rightarrow k + w$.

2. Return $\{z_k\}$ for $k = 1, 2, \dots, N$.

Note that taking $p_1 = q$ and $p_0 = 1 - q$ yields a Bernoulli sequence with probability q that $z_k = 1$; an example is shown in the upper plot in Fig. 2, which was obtained by taking $p_0 = 0.95$ and $p_1 = 0.05$. Conversely, taking $p_w > 0$ for $w > 1$ yields binary sequences with patches of width w . As a specific example, the sequence shown in the lower plot in Fig. 2 was obtained by setting $p_0 = 0.95$ and $p_3 = 0.05$, giving a sequence that always exhibits patches of length 3, an outcome that occurs 5% of the time. Note, however, that since each 1 generated by this model occurs three times in succession, the probability that $z_k = 1$ is 15% rather than 5% as in the Bernoulli example. More generally, note that the expected number of 1's in the binary sequence $\{z_k\}$ generated according to the procedure just described is

$$(3.3) \quad N_a = N \sum_{w=0}^N w p_w.$$

Hence, if we wish to generate a patchy sequence of fixed length N having a specified value for N_a (e.g., a patchy outlier sequence with "10% contamination"), it is necessary to reduce the probabilities p_w accordingly for $w > 0$ to account for the patch effects. This modification is equivalent to increasing the probability p_0 , an idea closely related to the use of zero-inflated Poisson models [4, 11], zero-inflated binomial models [11], or zero-inflated negative binomial models [4] in analyzing count data.

4 Empirical measures of patchiness

To assess the patchiness of a status sequence $\{z_k\}$ of length N , this paper considers three closely related empirical measures. The first is the *empirical concentration*

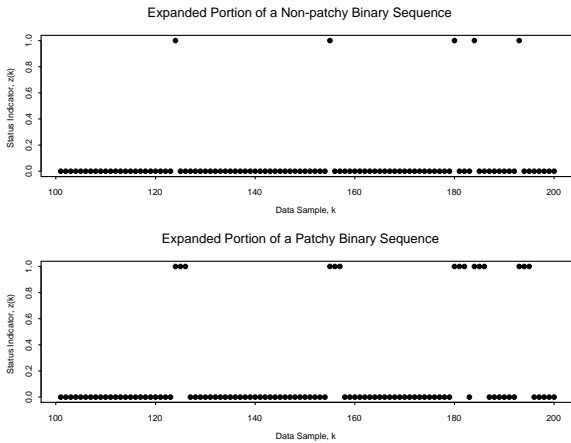


Figure 2: Two simulated binary sequences: a non-patchy Bernoulli sequence, assuming the value $z_k = 1$ with probability 5% (top), and a sequence that exhibits only patches of width 3 (bottom).

measure:

$$(4.4) \quad \hat{\phi}_w = \frac{N_w(w+1)}{N-1},$$

which represents the fraction of the maximum possible number of patches of width w , based on the following definitions. A *patch of width w* in the binary sequence $\{z_k\}$ is defined by the following two conditions, which must hold for some index k satisfying $k \geq 2$ and $k \leq N - w$:

$$(4.5) \quad \begin{aligned} z_k = z_{k+1} = \dots = z_{k+w-1} &= 1 \\ \text{and } z_{k-1} = z_{k+w} &= 0. \end{aligned}$$

It follows from this definition that the maximum possible number of patches satisfies:

$$(4.6) \quad N_w w \leq (N-2) - (N_w - 1).$$

The term on the left-hand side of this inequality counts the number of points included in the patches of width w , while the first term on the right-hand side is the maximum number of points in the sequence $\{z_k\}$ that can assume the value $z_k = 1$ (i.e., z_1 and z_N must both be zero), and the second term on the right-hand side is the minimum number of zero values required to separate successive patches of width w . It follows from Eq. (4.6) that

$$(4.7) \quad N_w \leq \frac{N-1}{w+1},$$

meaning that the empirical concentration measure $\hat{\phi}_w$ satisfies $0 \leq \hat{\phi}_w \leq 1$ for all patch widths w .

The numerical results presented in the following examples were computed in the *S-plus* software package;

while it is possible to compute $\hat{\phi}_w$ by brute force, this involves a very slow nested loop construction, and it is *much* faster to use an algorithm based on the following observations. First, define the complementary sequence $\{z_k^c\}$ to the status sequence $\{z_k\}$ by:

$$(4.8) \quad z_k^c = 1 - z_k.$$

Next, note that the two defining conditions given in Eq. (4.5) for a patch of width w may be expressed as:

$$(4.9) \quad \begin{aligned} z_{k-1}^c = z_k = z_{k+1} = \dots = z_{k+w-1} = z_{k+w}^c &= 1 \\ \Leftrightarrow z_{k-1}^c \cdot z_k \cdot z_{k+1} \cdot \dots \cdot z_{k+w-1} \cdot z_{k+w}^c &= 1. \end{aligned}$$

The advantage of this observation is that it immediately yields the following expression for the number N_w of patches of width w in the sequence $\{z_k\}$, i.e.

$$(4.10) \quad N_w = \sum_{k=2}^{N-w} z_{k-1}^c \cdot z_k \cdot z_{k+1} \cdot \dots \cdot z_{k+w-1} \cdot z_{k+w}^c.$$

To compute $\hat{\phi}_w$ over a range of values from 1 to some maximum patch width w^* , simply construct a matrix with N rows and w^* columns where each column contains the vector appearing on the right-hand side of Eq. (4.10). The number N_w can then be efficiently computed as the vector of row sums of this matrix and $\hat{\phi}_w$ can be computed directly from this result.

By itself, a sequence of values $\{\hat{\phi}_w\}$ computed from a given binary status sequence $\{z_k\}$ is not easy to interpret: does $\hat{\phi}_w = 0.3$ give strong or weak evidence in support of the hypothesis that $\{z_k\}$ exhibits an unusually large number of patches of width w ? Conversely, how small must $\hat{\phi}_w$ be to suggest *fewer* patches of width w than we would expect under the homogeneous Bernoulli alternative? In subsequent discussions, this latter phenomenon will be called *sparseness*. To address these questions, this paper adopts a permutation strategy [10], analogous to that used previously in assessing the significance of clustering results [17]. Specifically, given a sequence $\{z_k\}$, applying a random permutation to the index sequence should destroy any patchiness that may be present, effectively reducing the randomized sequence $\{\tilde{z}_k\}$ to a Bernoulli sequence. Hence, if the number of patches of width w is unusually large in the original sequence, relative to a Bernoulli alternative, this randomization should cause $\hat{\phi}_w$ to decrease significantly. Similarly, if the sequence $\{z_k\}$ contains significantly fewer patches of width w than expected under the Bernoulli alternative, randomization should cause $\hat{\phi}_w$ to increase. Repeating this process for M statistically independent random permutations gives a sequence $\{\tilde{\phi}_w^j\}$ of patchiness measures that can be used

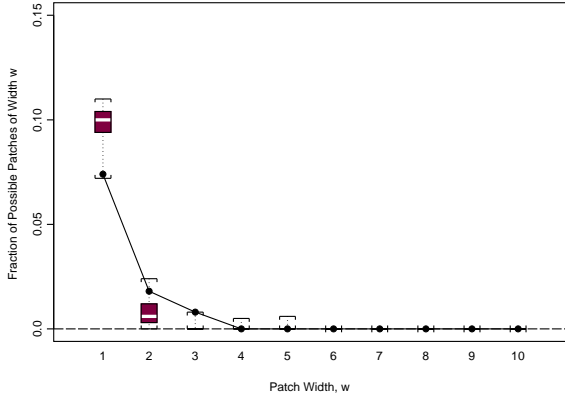


Figure 3: Empirical patchiness characterization for the Bernoulli sequence shown in the upper plot in Fig. 2. The line through the solid circles corresponds to the computed values of $\hat{\phi}_w$. The boxplots each summarize the $\hat{\phi}_w^j$ values obtained from 200 randomizations of the original sequence.

to assess the significance of the original patchiness measure $\hat{\phi}_w$. In particular, $\hat{\phi}_w$ gives evidence of patchiness if it lies above the range of the randomization results, and it gives evidence of sparseness if it lies below this range. Since these comparisons define a two-sided hypothesis test, $\hat{\phi}_w$ values falling outside this range are significant at the level $2/M$.

Fig. 3 summarizes the results obtained for the Bernoulli sequence shown in the upper plot in Fig. 2 using the empirical patchiness characterization proposed here, for patch widths between $w = 1$ and $w = 10$. The values of $\hat{\phi}_w$ computed from the original sequence $\{z_k\}$ are shown in Fig. 3 as solid circles, connected by a line, and the results $\{\hat{\phi}_w^j\}$ obtained for 200 independent randomizations are summarized with boxplots. Overall, the results are exactly what we expect for a Bernoulli sequence, which may be taken as a reference standard of “non-patchiness.” Specifically, these results show no evidence of patchiness since none of the $\hat{\phi}_w$ values fall outside the range of values generated by the 200 randomizations. Also, note that the quantity $\hat{\phi}_w$ shown here is not the same as the *patchy contamination at width w* $\hat{\gamma}_w$, defined as the number of contaminants contributed by patches of width w and given by

$$(4.11) \quad \hat{\gamma}_w = \frac{N_w w}{N} = \left(\frac{w}{w+1} \right) \left(\frac{N-1}{N} \right) \hat{\phi}_w.$$

In particular, note that $\hat{\gamma}_1 \simeq 0.05$, corresponding to the total contamination level of the Bernoulli sequence considered here, while $\hat{\phi}_1 \simeq 0.10$, twice this level.

For comparison, Fig. 4 shows the corresponding results obtained from the patchy sequence shown in the bottom plot in Fig. 2, which consists entirely of patches of width 3. As before, the original ϕ_w values are represented by the solid circles connected with the smooth curve, and the 200 randomization results $\{\tilde{\phi}_w^j\}$ are summarized with boxplots. Because of the special character of the status sequence $\{z_k\}$ considered here, these results illustrate both significant patchiness and significant sparseness, relative to the Bernoulli alternative. In particular, since no patches of widths 1 or 2 appear in this sequence, $\hat{\phi}_1 = \hat{\phi}_2 = 0$ here and these results fall well below the range of the corresponding randomization values. In other words, these results correctly reflect the extreme sparseness of the sequence $\{z_k\}$ with respect to patches of widths $w = 1$ and $w = 2$. Conversely, the results for $\hat{\phi}_3$ lie well above the range of the randomization results, giving strong evidence in support of the patchiness hypothesis for $w = 3$. None of the other results are significant, as they all fall within the range of the corresponding randomizations. Note, however, that although the value of $\hat{\phi}_6$ is not significant relative to the randomizations, it is the only nonzero result other than $\hat{\phi}_3$, reflecting the fact that two successive patches of width 3 were generated here by the random sequence generator described in Sec. 3 with no intervening zero value, converting them into a single patch of width 6.

The visual similarity of this result to a second harmonic in a power spectrum suggests the following alternative graphical representation of the results presented here. Define $\hat{\psi}_w$ as the *patch spectrum of width w*, given by

$$(4.12) \quad \begin{aligned} \hat{\psi}_w &= \frac{N_w w}{\sum_{k=1}^N z_k} \\ &= \left(\frac{(N-1)w}{(w+1) \sum_{k=1}^N z_k} \right) \hat{\phi}_w. \end{aligned}$$

Note that this quantity represents the fractional contribution of patches of width w to the total number of 1’s in the $\{z_k\}$ sequence. Because $\hat{\psi}_w$ is linearly related to $\hat{\phi}_w$ by a constant that is invariant under random permutations, it follows that the randomization results $\tilde{\psi}_w^j$ may be computed directly from the corresponding randomization results $\tilde{\phi}_w^j$, i.e.,

$$(4.13) \quad \tilde{\psi}_w^j = \left(\frac{(N-1)w}{(w+1) \sum_{k=1}^N z_k} \right) \tilde{\phi}_w^j.$$

An advantage of $\hat{\psi}_w$ over $\hat{\phi}_w$ is that, while both patchiness measures are normalized to lie in the unit interval $[0, 1]$, this upper limit has a more useful interpretation

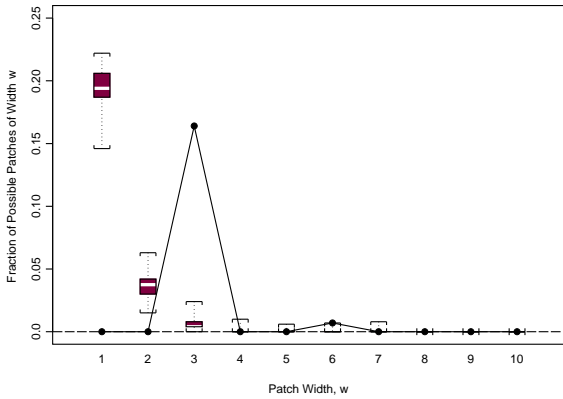


Figure 4: Empirical patchiness characterization $\hat{\phi}_w$ for the patchy binary sequence shown in the lower plot in Fig. 2, in the same format as Fig. 3.

for the patch spectrum $\hat{\psi}_w$ than it does for the empirical patchiness measure $\hat{\phi}_w$. For example, note that if $\hat{\phi}_1 = 1$, the sequence $\{z_k\}$ is completely specified: it is a periodic sequence of odd length N that takes the value $z_k = 0$ whenever k is odd and $z_k = 1$ whenever k is even. Although this sequence certainly can arise, it is not a typical status sequence we might expect to see in practice. Conversely, the result $\hat{\psi}_1 = 1$ is easily seen to arise if and only if every data anomaly is isolated, a much more likely situation in practice, and one that we might well be interested in detecting.

More generally, the scaling of the patch spectrum $\hat{\psi}_w$ appears to be much more informative than that of the empirical patchiness measure $\hat{\phi}_w$, as may be seen by comparing Figs. 4 and 5. In particular, the fact that $\hat{\psi}_3 \simeq 1$ demonstrates clearly that almost all of the 1's in the sequence $\{z_k\}$ appear in patches of width $w = 3$, a point that is not obvious from the numerical values of $\hat{\phi}_w$ plotted in Fig. 4. Similar conclusions apply for the Bernoulli sequence: the patch spectrum results in Fig. 6 show that $\sim 65\%$ of the 1's in this sequence occur in patches of width $w = 1$, $\sim 20\%$ in patches of width $w = 2$, and the remaining $\sim 15\%$ in patches of width $w = 3$. Again, this quantitative interpretation is not obvious from the numerical values for $\hat{\phi}_w$ shown in Fig. 3. Overall, because the results for $\hat{\psi}_w$ are easier to interpret than those for $\hat{\phi}_w$, the remainder of this paper focuses entirely on the patch spectrum $\hat{\psi}_w$.

5 Two numerical summary statistics

Although plots like Figs. 3 through 6 give useful qualitative characterizations of patchiness, it is desirable to also

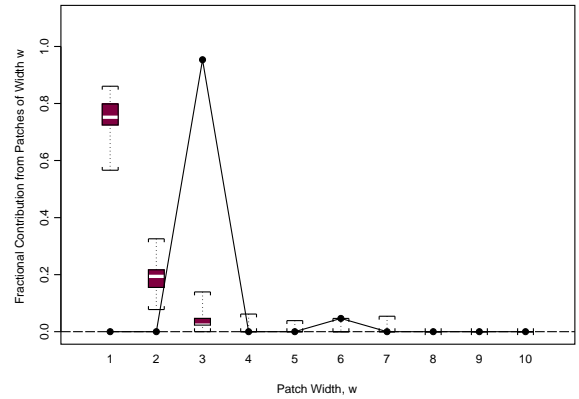


Figure 5: Computed patch spectrum $\hat{\psi}_w$ for the patchy binary sequence shown in the lower plot in Fig. 2, in the same format as Fig. 3.

have simple numerical summary statistics. The following discussion presents two such summaries: the z-score z_w of the patch spectrum $\hat{\psi}_w$ relative to the randomized results $\{\tilde{\psi}_w^j\}$, and a quantity α_w that gives a normalized measure of the distance $\hat{\psi}_w$ lies from the most extreme $\tilde{\psi}_w^j$ value, relative to the range of possible $\hat{\psi}_w$ values.

More specifically, let $\tilde{\mu}_w$ denote the mean of the M randomized values $\{\tilde{\psi}_w^j\}$ and let $\tilde{\sigma}_w$ denote the standard deviation of these values. The z-score for the $\hat{\psi}_w$ value computed from the original data sequence is then defined as

$$(5.14) \quad z_w = \frac{\hat{\psi}_w - \tilde{\mu}_w}{\tilde{\sigma}_w}.$$

If, as in the case of $\hat{\psi}_w$ for $w > p$, all of the randomized values $\tilde{\psi}_w^j$ are equal (e.g., zero in this case), it follows that $\tilde{\sigma}_w = 0$. When $\hat{\psi}_w$ also exhibits this common value, again as in the case of $\hat{\psi}_w$ for $w > p$, the value of z_w will be defined as zero; otherwise, z_w will be defined as $\pm\infty$, depending on the sign of $\hat{\psi}_w - \tilde{\mu}_w$.

If the randomized values $\{\tilde{\psi}_w^j\}$ exhibit an approximately normal distribution, we would expect to see no z-scores larger in magnitude than $|z_w| \sim 3$ for non-patchy status sequences $\{z_k\}$ (specifically, the probability of observing a normal random variable with a z-score of magnitude larger than 3 is approximately 0.3%). However, the shape of some of the boxplot summaries is strongly suggestive of significant asymmetry, bringing the appropriateness of approximate normality assumptions seriously into question. Still, it follows from Chebyshev's inequality [2, p. 75] that:

$$(5.15) \quad \mathcal{P} \left\{ \left| \frac{x - \mu}{\sigma} \right| > \beta \right\} = \mathcal{P} \{ |z| > \beta \} \leq \frac{1}{\beta^2},$$

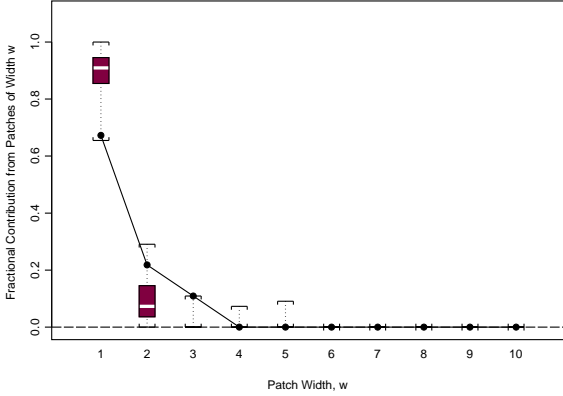


Figure 6: Computed patch spectrum $\hat{\psi}_w$ for the Bernoulli sequence shown in the upper plot in Fig. 2, in the same format as Fig. 3.

for any finite variance distribution. Hence, even under this very weak distributional assumption, it follows that z-scores of large magnitude are unlikely. In particular, note that $|z_w| > 10$ has probability less than 1% even under this extremely conservative working assumption.

The second summary statistic considered here is α_w , defined as follows. First, define the minimum and maximum random permutation values as:

$$(5.16) \quad \begin{aligned} \tilde{\psi}_w^- &= \min_j \{\tilde{\psi}_w^j\}, \\ \tilde{\psi}_w^+ &= \max_j \{\tilde{\psi}_w^j\}. \end{aligned}$$

The α_w value is then defined as:

$$(5.17) \quad \alpha_w = \begin{cases} \frac{\hat{\psi}_w - \tilde{\psi}_w^+}{1 - \tilde{\psi}_w^+} & \hat{\psi}_w > \tilde{\psi}_w^+ \\ 0 & \tilde{\psi}_w^- \leq \hat{\psi}_w \leq \tilde{\psi}_w^+ \\ -\left(\frac{\tilde{\psi}_w^- - \hat{\psi}_w}{\tilde{\psi}_w^-}\right) & \hat{\psi}_w < \tilde{\psi}_w^- \end{cases}$$

Note that α_w is nonzero if and only if $\hat{\psi}_w$ is significant with respect to the random permutation values $\{\tilde{\psi}_w^j\}$. For $\hat{\psi}_w$ values lying above the range of these permutation values, α_w is positive, bounded above by its maximum achievable value of 1. Since this behavior is precisely what we expect for a patchy status sequence, positive α_w values may be interpreted as a measure of the strength of evidence in support of the hypothesis that $\{z_k\}$ exhibits an unusually large number of patches of width w , relative to the Bernoulli model. In particular, note that $\alpha_w \simeq 1$ implies that essentially all of the anomalies identified by the status sequence $\{z_k\}$ occur in a patch of width w that has low probability under the Bernoulli model. Conversely, a sparse sequence will

exhibit *fewer* patches of width w than expected under the Bernoulli model, and this will give rise to negative α_w values, with the same general interpretation. Specifically, negative α_w values lie between -1 and 0 , and they may be viewed as a measure of the strength of evidence in support of the *sparseness hypothesis* that $\{z_k\}$ exhibits significantly fewer patches of width w than would be expected under the Bernoulli model. More specifically, $\alpha_w \simeq -1$ implies that patches of width w that are expected to be present under the Bernoulli model are largely absent in the observed status sequence $\{z_k\}$.

For the patchy status sequence $\{z_k\}$ shown in the bottom plot in Fig. 2, nonzero α_w values are observed only for $w = 1, 2,$ and 3 , and these values are $\alpha_1 = \alpha_2 = -1$ and $\alpha_3 = 0.95$. These results reflect first, the complete absence of patches of width $w = 1$ and $w = 2$, both expected under the Bernoulli alternative, and an overwhelming predominance of patches of width $w = 3$, which are relatively rare under the Bernoulli alternative. For comparison, none of the α_w values computed from the Bernoulli sequence shown in the upper plot in Fig. 2 are nonzero, in perfect agreement with our expectations.

It is particularly instructive to consider the z-scores for this example. For the patchy sequence shown in the bottom plot in Fig. 2, nonzero z-scores are obtained for all patch widths w between 1 and 7, although four of these values are quite small in magnitude (specifically, $z_4 = -0.4$, $z_5 = -0.1$, and $z_7 = -0.1$). The results for patch widths of 1 and 2 exhibit negative z-scores, consistent with the absence of expected patches of these widths in this example: $z_1 = -15.1$ and $z_2 = -4.3$. Not surprisingly, the results for patches of width 3 exhibit the largest magnitude z-score seen: $z_3 = 32.6$. What is somewhat surprising is that $z_6 = 9.9$, an extremely large z-score, especially for a result that is not significant with respect to the random permutations. Less extreme but somewhat similar results are obtained for the Bernoulli sequence shown in the top plot in Fig. 2: $z_1 = -3.6$, $z_2 = 2.2$, and $z_3 = 5.0$, with much smaller values for $w = 4$ and $w = 5$ ($z_4 = z_5 = -0.1$). The large magnitudes of some of these non-significant z-scores further emphasizes the point noted above that approximate normality should not be assumed for the permutation values $\{\tilde{\phi}_w^j\}$ here, since many of these z-scores would be extremely significant under the Gaussian model. As a practical matter, the best strategy is probably to consider only the z-scores for those values having nonzero α_w values.

6 Applications to the AERS database

The following examples serve both to illustrate the application of the patch spectrum $\hat{\psi}_w$ to real data, and to demonstrate that the characterization of patchiness

can lead to the identification of unusual structure even when applied to record indices (i.e., access keys) k with little or no inherent real-world significance. Both examples are based on the U.S. Food and Drug Administration’s Adverse Event Reporting System (AERS) database, which summarizes medical adverse events reported in conjunction with the use of specific drugs. A detailed description is available through the website <http://www.fda.gov/cder/aers/>. In general terms, exploratory analysis of this database is of interest because it can provide evidence of significant associations between specific drugs and adverse reactions, or between pairs of drugs. As a specific example, DuMouchel presents a Bayesian characterization of interesting drug-event combinations [9]. This database is examined here for two reasons: first, that it represents a real database of moderately large size as a practical testbed for the analysis methods described here and second, because unrecognized structure in a dataset can have a deleterious influence on analyses based on working assumptions that are inconsistent with this structure. As a specific example, Dodge [8] discusses the analysis of a dataset that has been adopted as a standard benchmark in the applied statistics literature (the Brownlee stack-loss dataset), noting that many authors have analyzed this dataset based on the assumption that the measurements represented a uniformly sampled time-series. He argues convincingly that this is not the case, bringing a number of these earlier conclusions (e.g., identification of specific observations as outliers) into question.

The AERS database is organized by quarter and year, and the specific data values used in both examples considered here were obtained from the the following five datasets from first quarter 2001 portion of this database: DEMO01Q1 gives demographic information, REAC01Q1 lists specific adverse reactions, DRUG01Q1 lists the drugs involved, OUTC01Q1 gives outcome information (e.g., “death,” “hospitalization,” or “other”), and RPSR01Q1 gives information pertaining to the source of each adverse event report. These files are linked via an integer primary key designated the ISR (Individual Safety Report) number for each record, which corresponds to a report logged by the FDA. For example, each record in the DEMO01Q1 dataset consists of the ISR, together with 13 other values, including the date the manufacturer first received information reported to the FDA, the name of the manufacturer sending the report, and the age and gender of the patient associated with the report. Similarly, each record in the REAC01Q1 data file consists of the ISR and a single character string describing a reported reaction. Since each report typically lists more than one adverse reaction, however, ISR’s generally appear more than

once in the REAC01Q1 dataset, unlike the DEMO01Q1 dataset, where each ISR appears only once. Analogous observations apply to the DRUG01Q1 and OUTC01Q1 data files: each adverse event report may have more than one entry.

6.1 Application 1: missing data As is often the case, the fraction of missing data in the datasets that make up the AERS database varies strongly between fields. For example, the file DEMO01Q1 contains 51,012 records, each corresponding to a unique ISR number. Since this field is used as the primary access key for matching records across the first quarter 2001 datasets, is extremely well-maintained, containing neither duplicate nor missing entries. Similarly, the FDA report date, the field indicating the date that the FDA was notified of the reported adverse event, is also free of missing data. In contrast, some of the other twelve fields exhibit significant fractions of missing data:

- patient age: $\sim 27.5\%$ missing,
- patient gender: $\sim 6.8\%$ missing,
- manufacturer reporting date: $\sim 8.4\%$ missing,
- date of adverse event: $\sim 28.7\%$ missing.

Further, certain subsets of the data may exhibit much higher missing data percentages. As a specific example, consider the set of records for which both the manufacturer reporting date and the date of adverse event are missing. While this situation only arises in $\sim 0.9\%$ of the total data records, within this subset of records, gender is also missing about 38% of the time, and age is also missing about 59% of the time.

To explore this case further, consider the following question: do these missing records occur at random throughout the DEMO01Q1 data file, or do they exhibit evidence of significant patchiness? One reason this question is interesting is that, if these particular incomplete records group together by ISR number, they may also exhibit other common characteristics that are of significantly greater interest. To consider this question, define the binary status sequence $\{z_k\}$ as:

$$z_k = \begin{cases} 1 & \text{if both event date and manufacturer date} \\ & \text{are missing,} \\ 0 & \text{otherwise.} \end{cases} \quad (6.18)$$

Values of $\hat{\psi}_w$ were computed for $w = 1$ to $w = 20$ and these results satisfied the normalization condition, indicating that all patches had been found. Comparing these results with the corresponding values $\tilde{\psi}_w^j$ for $M = 200$ random permutations show that too few patches of

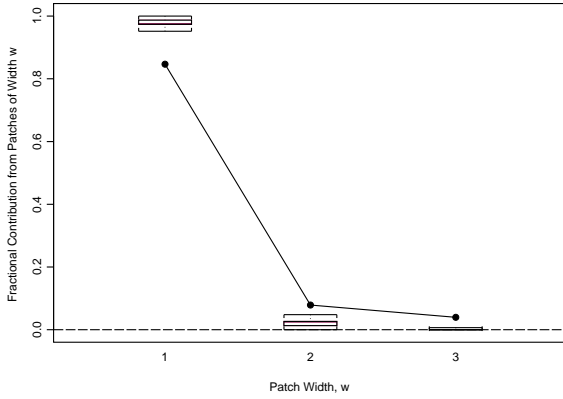


Figure 7: Computed patch spectrum $\hat{\psi}_w$ and $M = 200$ corresponding random permutation results for the AERS missing data status sequence $\{z_k\}$ for patch widths $w = 1, 2$, and 3 .

width $w = 1$ were seen in the sequence $\{z_k\}$ relative to the Bernoulli alternative, and too many patches of widths $2, 3, 4, 5$ and 7 were observed.

Fig. 7 shows these results for patch widths $w = 1, 2$ and 3 . Specifically, the solid circles in this plot represent the patch spectrum values $\hat{\psi}_w$ computed from the original status sequence for these values of w , while the boxplots summarize the range of $\hat{\psi}_w^j$ values obtained for the 200 random permutations considered here. It is clear from this plot that the value $\hat{\psi}_1$ lies well below the range of the randomizations, indicating as noted above that there are too few patches of width $w = 1$, relative to the Bernoulli alternative. Similarly, the values of $\hat{\psi}_2$ and $\hat{\psi}_3$ both lie above the range of the randomization values.

The patch spectrum results $\hat{\psi}_w$ and their associated randomizations $\hat{\psi}_w^j$ are summarized in Fig. 8 for patch widths $w = 3$ through $w = 10$. This plot was separated from Fig. 7 to emphasize small but significant details; in particular, note that the scale of Fig. 7 spans the range from 0 to 1 on the vertical axis, while Fig. 8 spans the narrower range from 0 to 0.05 . This plot emphasizes that, while patches of width $w = 3$ do occur in the randomizations, they are much less frequent than in the non-randomized results. In addition, it is clear that patches of width $w = 4, 5$, and 7 appear in the original sequence but *not* in the 200 randomizations.

It is instructive to examine the results for $w = 7$ in more detail, which corresponds to a single patch. Examination of these seven adverse event reports reveals that:

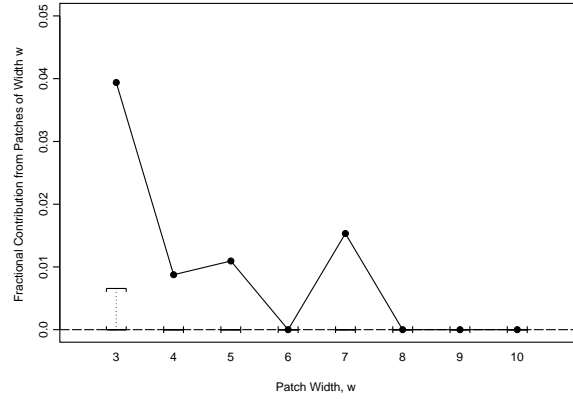


Figure 8: Computed patch spectrum $\hat{\psi}_w$ and $M = 200$ corresponding random permutation results for the AERS missing data status sequence $\{z_k\}$ for patch widths $w = 3$ through $w = 10$.

- the same FDA report date is listed for all ISR's,
- both age and gender are always missing,
- none of the fields in dataset DEMO01Q1 that identify manufacturer contain entries,
- none of these ISR's have a corresponding entry in the RPSR01Q1 report source information file,
- the ISR's all implicate different drugs,
- *all ISR's list the same, single reaction: "drug maladministration."*

The key points here are first, that this collection of seven successive records share many unusual characteristics in common. Hence, even though ISR number is a completely un-interesting data field by itself, patches of successive ISR's sharing a few anomalous characteristics (here, missing manufacturer report date and event date) actually share a much wider range of unusual characteristics. The second key point is that the patches detected by the method described here represent *extremely* small subsets of the data: in this example, a sequence of 7 anomalous records is detected, out of a total of $51,012$ (approximately 0.01%).

6.2 Application 2: death outcomes Of the $51,012$ adverse event reports with records in the DEMO01Q1 dataset, $5,048$ have "death" listed as an outcome in the OUTC01Q1 dataset. Here, we adopt this outcome as a subjective interestingness measure

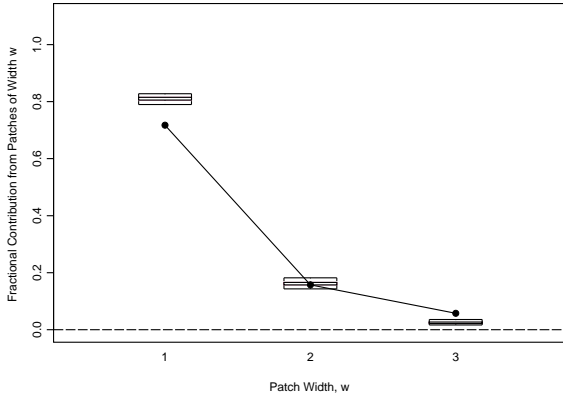


Figure 9: ISR patch spectrum $\hat{\psi}_w$ for patches of width $w = 1, 2$ or 3 for the 1Q 2001 AERS data listing “death” as the outcome.

[12] and consider the following status sequence:

$$(6.19) z_k = \begin{cases} 1 & \text{if ISR } k \text{ has outcome “death,”} \\ 0 & \text{otherwise.} \end{cases}$$

Fig. 9 shows the patch spectrum $\hat{\psi}_w$ values for $w = 1, 2,$ and 3 , denoted by solid circles connected by a line, with the results from 200 random permutations shown as boxplots. It is clear from this plot that isolated ISR’s (i.e., patches of width $w = 1$) occur less frequently than would be expected under the Bernoulli alternative, that patches of width $w = 2$ are consistent with what we would expect from a Bernoulli sequence, and patches of width $w = 3$ occur more frequently than we would expect for a Bernoulli sequence. Fig. 10 shows the corresponding results for w between 3 and 20, again plotted on a different scale to show the details more clearly. It may be seen from this figure that the status sequence $\{z_k\}$ defined in Eq. (6.19) exhibits an unusually large number of patches of widths 3 through 10. Even more unusual is the result for patches of width $w = 19$, which have essentially zero probability under the Bernoulli model.

A more complete quantitative summary of these results is given in Table 1, which lists the values computed for $\hat{\psi}_w$, the associated z-score z_w , the corresponding α_w value, and the number of patches N_w for all results giving nonzero $\hat{\psi}_w$ values between $w = 1$ and $w = 50$. Since these $\hat{\psi}_w$ values sum to 1, they account for all of the ISR’s listing “death” as their associated outcome. In fact, it follows from the results shown in Table 1 that approximately 12.5% of the death ISR’s occur in successive groupings of width 3 or more, with the most extreme one having a width of $w = 35$. Also, note that

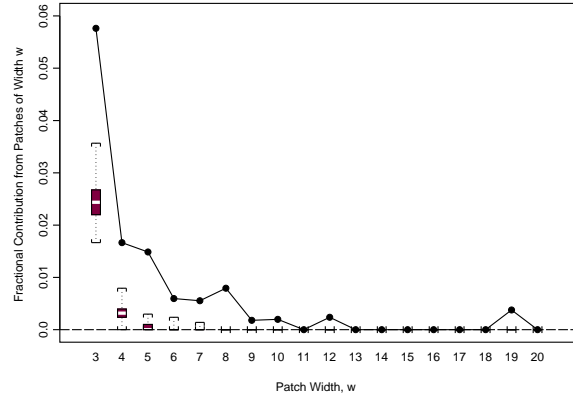


Figure 10: ISR patch spectrum $\hat{\psi}_w$ for patches of width $w = 3$ through $w = 20$ for the 1Q 2001 AERS data listing “death” as the outcome.

this example illustrates the separate utility of the z_w and α_w values. In particular, the fact that the α_w values are small but nonzero means that, while no single patch width $w > 2$ contributes a majority of interesting ISR’s, these patches are all significant relative to the Bernoulli alternative. In particular, it is not difficult to show that $\alpha_w \leq \hat{\psi}_w$ whenever $\alpha_w > 0$ and that this bound holds with equality if and only if $\tilde{\psi}_w^+ = 0$, further implying that $\tilde{\psi}_w^j = 0$ for all randomizations j . In this case it also follows that $\tilde{\sigma} = 0$, which is responsible for the infinite z-scores shown in Table 1 for $w = 8, 9, 10, 12, 19,$ and 35 since this condition holds for these cases.

It is particularly instructive to look at the most extreme case, $w = 35$. As indicated in Table 1, only a single patch of this width occurs, and examining the corresponding records from the DEMO01Q1 file reveal that 31 of these 35 successive ISR’s share a common reporting manufacturer (Company A). To examine this result further, we can adopt this reporting manufacturer as a measure of interestingness and repeat the previous analysis. Specifically, define the binary status sequence:

$$(6.20) z_k = \begin{cases} 1 & \text{if reported by Company A,} \\ 0 & \text{otherwise.} \end{cases}$$

A plot of the patch spectrum computed from this sequence is shown in Fig. 11 for $w = 1$ through $w = 10$. Although this plot does not show the peak at width $w = 31$ that led to the construction and examination of this status sequence, it is clear from Fig. 11 that the number of isolated ISR’s is vastly smaller than would be expected under the Bernoulli alternative, and that the number of patches of widths 2 through 8 is much larger than would be expected. Even more significantly,

w	$\hat{\psi}_w$	z_w	α_w	N_w
1	0.7173	-13.0	-0.0918	3621
2	0.1573	-0.7	0.0000	397
3	0.0576	9.5	0.0228	97
4	0.0166	9.0	0.0088	21
5	0.0149	24.0	0.0119	15
6	0.0059	22.5	0.0036	5
7	0.0055	56.5	0.0042	4
8	0.0079	$+\infty$	0.0079	5
9	0.0018	$+\infty$	0.0018	1
10	0.0020	$+\infty$	0.0020	1
12	0.0024	$+\infty$	0.0024	1
19	0.0038	$+\infty$	0.0038	1
35	0.0069	$+\infty$	0.0069	1

Table 1: Patch spectrum $\hat{\psi}_w$, z-scores z_w , α_w values and number of patches of width w present in the status sequence for the AERS ISR’s with outcome “death.”

the single patch of width 31 in the Company A status sequence—the only patch of width wider than 10 in this sequence—accounts for approximately 26% of the total number of ISR’s associated with this manufacturer.

Further examination of the Company A results reveals the following details. Altogether, this company appears as reporting manufacturer in 118 ISR’s. Of these, 111 list “death” as an outcome, with the following additional characteristics:

- patient gender is missing in all ISR’s,
- one or both of the following reactions is listed for every ISR: “Non-Accidental Overdose,” or “Overdose Nos (Not Otherwise Specified),”
- the same manufacturer date is listed, corresponding to the date the manufacturer initially received notification of the adverse event.

In view of these results, it seems likely that the patchiness seen in these ISR sequence is due to the manner in which these adverse events were reported and processed by the FDA. Despite the fact that this patch generation mechanism is not especially interesting, it has led us to focus on a very interesting group of ISR’s. In particular, the results presented here demonstrate that the detection and interpretation of patchiness in sequences of data records can ultimately lead us to groups

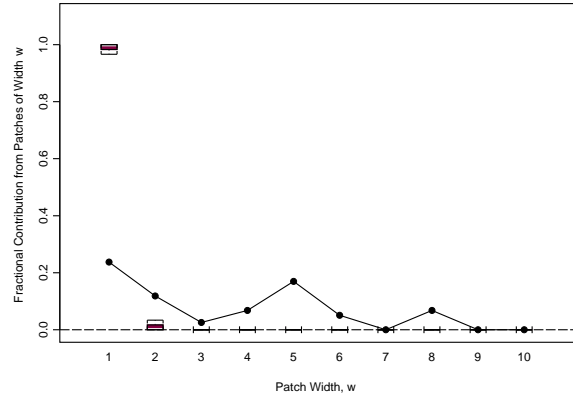


Figure 11: Patch spectrum for the Company A binary sequence defined in Eq. (6.20).

of records that are very strongly associated by characteristics that may be of significant interest (e.g., death by drug overdose).

7 Summary

This paper has considered the problem of detecting, characterizing, and interpreting non-random membership patterns of some class \mathcal{K} of data records in a larger dataset. Specific examples include missing or incomplete records, various types of data anomalies (e.g., outliers, inliers, or near-duplicate records), or record classes selected by either objective or subjective interestingness measures. The fundamental basis for these results is the binary status sequence $\{z_k\}$ defined in Eq. (1.2) indicating whether the record \mathcal{R}_k belongs to class \mathcal{K} or not. The main tools introduced here to characterize this sequence are the patch spectrum $\hat{\psi}_w$ introduced in Sec. 4 and the associated summary statistics z_w and α_w introduced in Sec. 5. The effectiveness of these tools was demonstrated first for a pair of simulation-based examples (a non-patchy Bernoulli sequence and a sequence exhibiting only random patches of width $w = 3$), and then with respect to status sequences constructed from the FDA’s AERS adverse event database. These examples illustrate that, even though the patches appearing in these record sequences are almost certainly data entry artifacts, they are interesting because the records involved were processed together for a reason related to some underlying common structure. For example, observation that 31 of 35 records in the longest observed patch of successive ISR’s with “death” listed as an outcome were associated with the same manufacturer led to an examination of all ISR’s associated with this man-

ufacturer. In turn, this led to the discovery that 111 of the 118 reports associated with this manufacturer were fatal drug overdoses, all with the same reporting date.

These examples also illustrated that, even in very long binary status sequences, the observation of wide patches is rare enough that they are easy to detect even if they represent an extremely small portion of the total sequence. For example, the AERS sequences considered here were of length 51,012, but the analysis methods presented here had no difficulty detecting single patches of width ~ 10 as unexpected features in the data, even though they correspond to $\sim 0.02\%$ of the data. As a corollary, this observation means that even if some unusual event or data recording anomaly places a single small patch of interesting records together, the fact that they are grouped together greatly enhances our ability to detect them. As a practical matter, even if this grouping is primarily due to the details of the data entry procedure, the fact that a group of records sharing the same characteristic of interest were entered together usually means that these records share several characteristics in common, as in the examples considered here. The results presented here suggest that patchiness analysis may be a very useful first step in uncovering these associations.

Finally, note that once we have constructed the binary status sequence $\{z_k\}$, we can apply a range of standard binary data analysis methods like logistic regression to explore possible relationships with other variables [5]. Alternatively, since $\{z_k\}$ defines a binary classification of records, we can also adopt the methodology of *case-control studies* [5, p. 217] or *case-referent studies* [19, p. 7]. There, the idea is to match each member of the “interesting class” (i.e., each record \mathcal{R}_k with $z_k = 1$) to one or more records from the “nominal class” (i.e., records \mathcal{R}_k with $z_k = 0$), usually subject to an approximate matching constraint on other record characteristics. The objective of these studies is to identify systematic differences in other characteristics that may be responsible for the difference in interestingness.

References

- [1] J. Albert and P. Williamson, “Using Model/Data Simulations to Detect Streakiness,” *Amer. Statistician*, vol. 55, 2001, pp. 41–50.
- [2] P. Billingsley, *Probability and Measure*, 2nd ed., Wiley, 1986.
- [3] A.Z. Broder, “Identifying and Filtering Near-Duplicate Documents,” in *Combinatorial Pattern Matching*, R. Giancarlo and D. Sankoff, eds., Springer-Verlag, 2000, pp. 1–10.
- [4] Y.B. Cheung, “Zero-inflated models for regression analysis of count data: a study of growth and development,” *Statist. Med.*, v. 21, 2002, pp. 1461–1469.
- [5] D. Collett, *Modelling Binary Data*, 2nd ed., Chapman and Hall, 2003.
- [6] L. Davies and U. Gather, “The identification of multiple outliers,” *J. Amer. Statist. Assoc.*, v. 88, 1993, pp. 782–801.
- [7] D. DesJardins, “Outliers, Inliers and Just Plain Liars—New Graphical EDA+ (EDA Plus) Techniques for Understanding Data,” *CEASAR Conf. Proc. Statistics Italy*, Rome, 2001, paper no. 169–26.
- [8] Y. Dodge, “The Guinea Pig of Multiple Regression,” in *Robust Statistics, Data Analysis, and Computer Intensive Methods*, H. Rieder, ed., Springer-Verlag, 1986, pp. 91–117.
- [9] W. DuMouchel, “Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System (with discussion),” *American Statistician*, v. 53, 1999, pp. 177–202.
- [10] P. Good, *Permutation Tests*, Springer-Verlag, 2000.
- [11] D.B. Hall and K.S. Berenhaut, “Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models,” *Canadian J. Statistics*, v. 30, 2002, pp. 1–16.
- [12] R.J. Hilderman and H.J. Hamilton, “Evaluation of Interestingness Measures for Ranking Discovered Knowledge,” *Proc. 5th Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2001, pp. 247–259.
- [13] K.-M. Lee, P. Meer, and R.-H. Park, “Robust Adaptive Segmentation of Range Images,” *IEEE Trans. Pattern Analysis Machine Intelligence*, v. 20, 1998, pp. 200–205.
- [14] R.D. Martin and V.J. Yohai, “Influence functionals for time-series,” *Ann. Statist.*, vol. 14, 1986, pp. 781–785.
- [15] R.K. Pearson, *Discrete-Time Dynamic Models*, Oxford, 1999.
- [16] R.K. Pearson, “Outliers in Process Modelling and Identification,” *IEEE Trans. Control Systems Technology*, v. 10, 2001, pp. 55–63.
- [17] R.K. Pearson, T. Zylkin, J.S. Schwaber, and G.E. Gonye, “Quantitative Evaluation of Clustering Results Using Computational Negative Controls,” *Proc. 2004 SIAM International Conference on Data Mining*, April, 2004, Lake Buena Vista, Florida, pp. 188–199.
- [18] D. Pyle, *Data Preparation for Data Mining*, Academic Press, 1999.
- [19] P.R. Rosenbaum, *Observational Studies*, 2nd ed., Springer-Verlag, 2002.
- [20] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, 1987.
- [21] W.E. Winkler, “Problems with Inliers,” paper presented at the European Conference of Statisticians, Prague, October, 1997.
- [22] Q. Zheng, K. Xu, W. Lv, and S. Ma, “Intelligent Search of Correlated Alarms from Database Containing Noise Data,” *Proc. 8th IEEE/IFIP Network and Operations Management Symposium (NOMS)*, Florence, Italy, 2002, pp. 405–419.