Fieller Stability Measure: A Novel Model-dependent Backtesting Approach*

Cristián Bravo†, Sebastián Maldonado‡

Abstract

Dataset shift is present in almost all real-world applications, since most of them are constantly dealing with changing environments. Detecting fractures in datasets on time allows recalibrating the models before a significant decrease in the model's performance is observed. Since small changes are normal in most applications and do not justify the efforts that a model recalibration requires, we are only interested in identifying those changes that are critical for the correct functioning of the model. In this work we propose a model-dependent backtesting strategy designed to identify significant changes in the covariates, relating a confidence zone of the change to a maximal deviance measure obtained from the coefficients of the model. Using logistic regression as a predictive approach, we performed experiments on simulated data, and on a real-world credit scoring dataset. The results show that the proposed method has better performance than traditional approaches, consistently identifying major changes in variables while taking into account important characteristics of the problem, such as sample sizes and variances, and uncertainty in the coefficients.

Keywords: Concept drift, Dataset shift, Backtesting, Credit scoring, Statistics, Logistic Regression.

*This is a post-peer-review, pre-copyedit version of an article published in the Journal of the Operational Research Society. The definitive publisher-authenticated version should be cited as "Cristián Bravo, Sebastián Maldonado (2015). 'Fieller Stability Measure: a novel model-dependent backtesting approach'. Journal of the Operational Research Society 66: 11. 1895 - 1905.", doi:10.1057/jors.2015.18, is available online at: `http://link.springer.com/article/10.1057\%2Fjors.2015.18`

†Departamento de Ingeniería Industrial, Universidad de Talca. Camino a Los Niches Km. 1, Curicó, Chile.

‡Universidad de Los Andes, Mons. Álvaro del Portillo 12455, Las Condes, Santiago, Chile.

Introduction

Dataset shift has received growing attention during the previous decade (Quiñonero Candela *et al.*, 2009; Moreno-Torres *et al.*, 2012). Most real-world domains deal with the issue that major changes may occur after the development of a predictive model, and the new data may not agree with the model which was estimated from past instances (Quiñonero Candela *et al.*, 2009). Since data mining models assume that historical data work as adequate predictors of what will happen in the future, detecting important changes and "setting the alarms" on time reduces the risk of making important decisions based on the wrong predicting models, incurring high misclassification costs (Hofer and Krempl, 2013; Kelly *et al.*, 1999). The process of monitoring and detecting dataset shift is known in the OR literature as *backtesting*, and represents one of the important challenges in data mining and business analytics for the upcoming years (Baesens *et al.*, 2009).

One particular example of this is credit scoring, defined as the estimation of the probability of default (i.e. the event of a borrower not repaying a loan in a given time period) by using statistical models to transform relevant data into numerical measures that guide credit decisions (Anderson, 2007). Adjustments in credit decisions are common since they are strongly affected by changes in the economy, political decisions, and regulation, causing different sample biases and leading to important changes in the customer portfolio. A credit scoring model that loses its predictive capacity may lead to significant costs: granting loans with high risk on one side, and denying low risk loans (which translates to an opportunity cost) on the other (Castermans *et al.*, 2010).

The aim of this paper is to provide a novel backtesting methodology for detecting concept drift. Our proposal relates two measures: one obtained from the coefficients of the original model used as a confidence zone for maximum covariate deviance, and the second one based on the ratio of the covariate means of the original and the new sample. While the first measure is obtained in a straightforward manner since the coefficients are approximately normally distributed, thanks to the Central Limit Theorem, the second measure provides a range for the actual change in the covariates in the new sample and it is obtained by using the Fieller's theorem (Fieller, 1954). The main advantages of our proposal compared with existing approaches can be summarized in three main points: our measure uses the model coefficients to construct a confidence zone that allows the detection of covariate shift, relating concept drift with the

estimation process. Second, our approach is suitable for any generalized linear model, for which confidence zones related to each variable can be obtained from the original model. Finally, our measure does not require the labels of the new samples, providing alerts before having evidence that the model is losing performance.

This paper is structured as follows: Recent developments in dataset shift are reviewed in the next section. The proposed concept shift approach is presented in the section that follows. In section Experimental Results we show the effectiveness of the presented approach using simulated data and a real-world credit scoring dataset. The main conclusions can be found in the final section of this paper, together with future developments we propose based on this research.

Prior Work in Dataset Shift and Backtesting

The term *dataset shift*, was first used by Quiñonero Candela *et al.* (2009) and was defined as "cases where the joint distribution of inputs and outputs differs between training and test stages". It has been studied during the last few years, and appears in the literature under different terms, including *concept/population drift* (Kelly *et al.*, 1999; Schlimmer and Granger, 1986), *contrast/shift mining* (Hofer and Krempl, 2013; Yang *et al.*, 2008), *change/fracture detection* (Cieslak and Chawla, 2007) and *backtesting* (Castermans *et al.*, 2010; Lima *et al.*, 2011), among others. A unifying vision of the concept was stated in Moreno-Torres *et al.* (2012). According to this work, dataset shift can be divided into three types: *covariate or population shift*, which refers to changes in the distribution of the input variables, generally studying shifts in the mean (Robinson *et al.*, 2002); *prior probability shift*, which occurs when the distribution of the class labels changes over time; and *concept shift*, when the relationship between the input variables and labels changes. Our proposal studies covariate shift, which is the most frequently studied form of dataset shift (Quiñonero Candela *et al.*, 2009). Prior probability shift and concept shift require labels for the new samples, which in some applications (such as application scoring) may take several months to become available. Since covariate shift can cause concept shift, it can also be used also for early detection of other forms of dataset shift (Moreno-Torres *et al.*, 2012). The advantage of our proposal compared to other covariate shift approaches is that additional information is incorporated, such as the stability of the original model, given by the confidence interval of its coefficients.

Several backtesting approaches have been proposed for dataset shift. We will discuss various models suggested in the literature next, which are also discussed as benchmark approaches in the relevant section. We use the following notation for a generic variable in a regression model: we define X_o as the original sample, or the one used to construct the predictive model, and X_n as the new sample, with elements $x_{io} \in X_o$, $i = 1, ..., |X_o|$, and $x_{i'n} \in X_n$, $i' = 1, ..., |X_n|$, where $|X_o|$ and $|X_n|$ represent the sample size of X_o and X_n, respectively. Vectors X_o and X_n are assumed to be numerical attributes with no missing values, which are further binned for measures Stability Index and Hellinger Distance. These metrics also allow nominal variables.

Stability Index

The Stability Index (SI) measures the similarity between two samples (Castermans *et al.*, 2010). For a given variable, this index can be computed as follows:

$$SI = \sum_{k=1}^{p} \left(\frac{X_{ko}}{|X_o|} - \frac{X_{kn}}{|X_n|} \right) ln \frac{X_{ko} \Big/ |X_o|}{X_{kn} \Big/ |X_n|}, \tag{1}$$

where $\frac{X_{ko}}{|X_o|}$ and $\frac{X_{kn}}{|X_n|}$ are the percentages of elements that belong to bin k for the reference and new sample, respectively. Both X_o and X_n contain p bins, in which each bin contains the count of some logical subunits measured between X_o and X_n.

Significant shifts in the population are indicated with higher values of the Stability Index, according to the following rules of thumb (Castermans *et al.*, 2010): An SI below 0.1 indicates no significant shift, an SI value between 0.1 and 0.25 means a minor shift, and an SI above 0.25 indicates a major shift.

The SI is intimately related to the concept of entropy, and, in particular, to the Kullback-Leibler divergence (D_{KL}) (Castermans *et al.*, 2010). For samples X_o and X_n, the Kullback-Leibler divergence considering each sample as a reference follows:

$$
\begin{aligned}
D_{KL}(X_o||X_n) &= \sum_{k=1}^{p} \frac{X_{ko}}{|X_o|} ln \frac{X_{ko} \Big/ |X_o|}{X_{kn} \Big/ |X_n|} \\
D_{KL}(X_n||X_o) &= \sum_{k=1}^{p} \frac{X_{kn}}{|X_n|} ln \frac{X_{kn} \Big/ |X_n|}{X_{ko} \Big/ |X_o|} = -\sum_{k=1}^{p} \frac{X_{kn}}{|X_n|} ln \frac{X_{ko} \Big/ |X_o|}{X_{kn} \Big/ |X_n|}
\end{aligned}
\tag{2}
$$

Therefore the Stability Index corresponds to the sum of both divergences, or a measure of the amount of incremental information required to recreate X_o from X_n and vice versa.

The Hellinger Distance

The Hellinger Distance indicates distributional divergence as a measure of similarity between two probability distributions (Basu *et al.*, 1997). Although the original Hellinger Distance is suitable for continuous variables, in our work we focus on the version used for backtesting in Cieslak and Chawla (2007), which requires either a categorical variable, or a previous binning step, similar to the Stability Index. The Hellinger Distance between the two samples follows:

$$
Hellinger(X_o, X_n) = \sqrt{\sum_{k=1}^{p} \left(\sqrt{\frac{X_{ko}}{|X_o|}} - \sqrt{\frac{X_{kn}}{|X_n|}} \right)^2} \tag{3}
$$

This measure reaches its minimum value (zero) when the distributions are identical, and its maximum ($\sqrt{2}$) when the distributions are completely divergent (i.e. no overlap for all bins).

The Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov (KS) Test is a non-parametric method that studies the divergence between two observed distributions under the hypothesis of independence (Smirnov, 1948). This test computes the maximum vertical deviation between the two empirical (cumulative) distribution functions, checking whether the two data samples come from the same distribution. The KS statistic for the samples X_o and X_n follows:

$$
D = \sup_x |F_o(x) - F_n(x)|, \tag{4}
$$

where $F_o(x)$ and $F_n(x)$ are the empirical cumulative distribution functions of X_o and X_n respectively. The KS test can be used as an alternative for the Student's t test, having the advantage that it makes no assumptions on the underlying distribution of the variables. This approach has been proposed for dataset shift detection in Cieslak and Chawla (2007). The same authors discuss the use of Pearson's χ^2 independence test as an alternative or complement for categorical variables, or previously binned ones.

Proposed Strategy for Dataset Shift

In this section we propose a model-dependent strategy for backtesting. The main objective of this method is to assess whether or not a model recalibration is necessary by identifying major changes in the distribution. Since traditional two-sample independence tests (Kolmogorov-Smirnov or χ^2 test, for instance) reject the hypothesis of dataset shift for small changes in the variables, as will be shown in section , we propose a measure that attempts to detect only critical changes in the distributions of the variables. The intuition behind our approach is that those changes occur when the difference between the observed and the reference sample exceeds a maximum deviance measure, which we obtain by computing a confidence zone based on the model coefficients estimated from the reference sample. Notice that this measure does not require labels for the observed sample, which is useful in applications in which the objective variable for any new sample is revealed only after an additional period of time. It is also important to highlight that the goal of this work is to develop a new metric for dataset shift than to define a formal statistical test for this purpose, since they have proven frail for this particular task.

The first step is to define when a shift in the variables is critical, in a way that determines if the current model is invalid for the new sample. Considering the coefficients $\boldsymbol{\beta}$ from a general linear model constructed from the reference sample X_o, we use the property that these estimators are asymptotically normal with a mean $\hat{\beta}$ standard error $se(\hat{\beta})$. Notice that this property is satisfied only for large sample sizes, although recent studies suggest that a sample size of 250 is sufficient to avoid the small sample bias (Bergtold *et al.*, 2011), and credit scoring models are seldom estimated with such a small portfolio. Subsequently, a parameter β has a confidence interval $[\beta^l, \beta^u]$, where the value of the parameter is found with a confidence level α. Using $z_\alpha = \Phi^{-1}(1 - \alpha/2)$, this interval can

be rewritten as follows:

$$[\beta^l, \beta^u] = [\hat{\beta} - z_\alpha se(\hat{\beta}), \hat{\beta} + z_\alpha se(\hat{\beta})] \tag{5}$$

We propose to use this confidence interval as a measure for maximum covariate deviance. This measure is a range within which we are comfortable with the changes observed in the new sample. It is important to notice that this confidence zone is not a hard boundary for the range of the covariates in the new sample.

We rescale the confidence zone by dividing it by $\hat{\beta}$, resulting in an interval composed of dimensionless elements (without any unit). We propose the following measure (Dimensionless Confidence Interval for β, DCI_β):

$$DCI_\beta \in [\frac{\hat{\beta} - z_\alpha se(\hat{\beta})}{\hat{\beta}}, \frac{\hat{\beta} + z_\alpha se(\hat{\beta})}{\hat{\beta}}] = [DCI_\beta^l, DCI_\beta^u] \tag{6}$$

The thinking behind this measure is that a shift in the covariates (in a relative, or percentual, measure) can be considered significant when the new data X_n is outside the confidence interval constructed with the reference data X_o, given by the coefficient obtained with the linear model for a particular variable and its standard deviation.

We now propose a second measure to assess the shift between the old and the new sample. Subsequently, we will relate this metric with DCI_β. Following the proposed notation, the mean of both samples can be computed as:

$$\bar{x}_o = \sum_{i=1}^{|X_o|} \frac{x_{io}}{|X_o|} \to \mu_o, \ \bar{x}_n = \sum_{i=1}^{|X_n|} \frac{x_{in}}{|X_n|} \to \mu_n, \tag{7}$$

where μ_o and μ_n are the population means of the original and the new variables, respectively, which we are assuming to be numerical attributes. We are interested in studying the behavior of the ratio of the two means (MR):

$$MR = \frac{\mu_n}{\mu_o} \tag{8}$$

The ratio of the means allows us to relate the two relevant pieces of information in our work: the relationship between the distributions of X_o and X_n, and the stability of the coefficients obtained from the original model. If there is no uncertainty in these parameters, then our hypothesis becomes MR=1, or, equivalently, that X_o and X_n have similar means. If this hypothesis does not

hold, we can set the alarms for dataset shift. Such a test can be performed with a standard two-sample t-test or a non-parametrical alternative, such as Kolmogorov-Smirnov or the Mann-Whitney U test. The intuition behind our approach is that the uncertainty of the coefficients demands stronger evidence of a dataset shift in order to set the alarms, and therefore we use the DCI as the hypothesis instead of MR=1. The use of a ratio for the means of the two samples and for the original coefficients allows us to construct unitless measures, making them comparable.

Since both means are population parameters, there is no certainty on the true value of the proportion of both measures. A procedure for estimating an approximate confidence interval for the proportion of two general means was developed by Fieller (1954). This confidence interval is computed as follows:

$$\begin{aligned}(FSM_l, FSM_u) &= \frac{1}{1-g}\left[\frac{\bar{x}_n}{\bar{x}_o} \pm \frac{t_{r,\alpha}}{\bar{x}_o}\left(S_n^2(1-g) + \frac{\bar{x}_n^2}{\bar{x}_o^2}S_o^2\right)^{1/2}\right] \\ g &= \frac{t_{r,\alpha}^2 S_n^2}{\bar{x}_o^2}\end{aligned} \tag{9}$$

in which FSM_l and FSM_u are the lower and upper bounds respectively of the Fieller Stability Measure (FSM); $t_{r,\alpha}$ is the critical value of a t-student with r degrees of freedom at an α level of confidence (for our experiments we use $\alpha = 5\%$). The degrees of freedom are computed as $r = a + b - 2$, where a is the sample size of the numerator and b is the sample size of the denominator. Assuming more than 30 loans in the sample, then $r > 30$, and therefore we can approximate $t_{r,\alpha} = z_\alpha = 1.96$. The unbiased estimators of the standard deviations for the original and new samples are S_o and S_n. Regarding potential correlation between the samples X_o and X_n, we assume that they are independent. The reason for this is that we assume that they are different elements (different loans) given that they are measured at different points in time. For application scoring (loans to borrowers who are first time customers of the company) this is clearly the case. For behavioral scoring this is not always so, but it has been established that the impact of having a lower number of cases that are equal in both samples is very minor in the final estimated parameters. Birón and Bravo (2014) discuss the effects of sample bias in the beta parameters in depth.

The original proportion by Fieller also allows correlation between the sam-

ples, but since both X_o and X_n are designed to be measured at different points in time, we can be certain that the two are independent.

The Fieller test is widely used in medicine (Beyene and Moineddin, 2005, see e.g.), measuring the cost-effectiveness of medical treatments by comparing the total medical cost of populations with and without a particular treatment, but, to the best of our knowledge, it has not been used to date as a measure of dataset shift.

We now want to relate both confidence zones: DCI_β from equation (6), and FSM from equation (9). Since dataset shift occurs when the mean of the new sample differs significantly from the reference sample, the DCI_β can be used as a measure of the maximum allowed deviation between samples. If the ratio of the two means falls outside this confidence zone would give a strong indicator that the sample has drifted beyond what can be seen as acceptable. This occurs if both confidence zones do not intersect, or, formally,

$$(FSM_l, FSM_u) \cap (DCI_\beta^l, DCI_\beta^u) = \emptyset, \tag{10}$$

then the shift between both samples is significantly larger than the maximum covariate deviance allowed by DCI_β, and then an alarm is set for this variable. We propose three different values for z_α, reflecting three levels of dataset shift: $z_\alpha = 1$ (DCI_β constructed based on one standard deviation over the mean), $z_\alpha = 2$ (DCI_β constructed based on two standard deviations over the mean), and $z_\alpha = 3$ (DCI_β constructed based on three standard deviations over the mean). If an alarm is set using $z_\alpha = 1$, then a *slight shift* is detected, where we recommend to monitor the variable in case of this shift sustains over time. Subsequently, if an alarm is set using $z_\alpha = 2$, a *relevant shift* is detected, and we suggest a model recalibration in order to avoid a significant loss of predictive power. Finally, if an alarm is set using $z_\alpha = 3$, a *severe shift* is detected, a recalibration is strongly recommended since this shift could lead to a significant degradation of the model's predictive capabilities and subsequent financial losses. This strategy is inspired by the Empirical Rule (Utts and Heckard, 2012), also known as the three-sigma rule, that states that nearly all values lie within three standard deviations of the mean in a normal distribution. The confidence levels for $z_\alpha = 1$, $z_\alpha = 2$, and $z_\alpha = 3$ are 31.7%, 4.5%, and 0.27%, respectively. Since our main goal is to create a stability measure rather than to define a formal test, note that α cannot be interpreted as error rates and are simply a natural parametrization for the uncertainty.

The proposed shift strategy has several advantages compared with existing dataset shift measures:

- It provides a nonarbitrary measure to assess significant changes in the covariates, on a statistical basis. This is an important advantage compared with methods like the Stability Index, in which the decision thresholds are tabulated according to arbitrary rules.
- Only major changes are highlighted using this strategy, compared with traditional two-sample tests like the Kolmogorov-Smirnov test, or χ^2 test, which tend to reject the independence hypothesis as soon as minor changes in the distribution occur.
- It takes into account of several properties of the data, such as the original model and the confidence in the estimated coefficients, the first two moments of the new and reference sample, and their respective sample sizes.

The comparison between the two confidence zones cannot be used to construct a statistical test, so our proposal has to be considered as a measure akin to the Stability Index. The reason for this is that the only conclusion that can be drawn from this comparison occurs when two confidence intervals do not overlap at all: one can be certain that the two elements are significantly different (which can be referred to as severe dataset shift), whereas when they do overlap, there is neither certainty nor uncertainty of the difference, and no 95% confidence statistics can be built. This fact is discussed in depth by Schenker and Gentleman (2001).

Experimental Results

In this section we compare the proposed approach with two well-known backtesting measures to demonstrate its ineffectiveness at identifying dataset shift: the Stability Index, which is of prime interest in credit scoring (Baesens, 2014), and the Kolmogorov-Smirnov test.

For all our experiments we used binary logistic regression on the original sample, since it is a well-known model with normal asymptotic properties (Hosmer and Lemeshow, 2000). It considers V different covariates in a vector $\mathbf{X}_i \in \mathbb{R}^V, i = 1, ..., |Xo|$, with labels $y_i = \{0, 1\}$. The probability of being from class 1 for a given sample i is:

$$p(y_i = 1|\mathbf{X}_i) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{V} \beta_j X_{ij}\right)}}, \tag{11}$$

where β_0 is the intercept, and β_j is the regression coefficient associated with variable j.

We first present empirical results using simulated data, and then describe experiments on a real-world credit scoring dataset.

Experiments on Synthetic Data

In this section we report how the proposed test behaves in a controlled experiment, isolating the effect of concept drift in the sample. To do so, we construct two synthetic datasets: a small one, which can be considered to be less stable, and a large dataset that has greater certainty on the parameters (and as such is more prone to dataset shift). Taking into account the original concurring advice given by Lewis (1992) and Siddiqi (2006), 1500 instances of each class should be sufficient to build robust models of high quality. The first dataset contains 1000 cases, below the recommended number and as such is assumed to be unstable, and the second has 5000 instances, with much greater certainty.

For each dataset, considering a class balance of 50% to eliminate any imbalance effects on the test, we require two distributions for the two subpopulations, induced by different values of the objective variable y. For $y = 1$, we generate 500 cases for the small synthetic dataset and 2500 cases for the large one at random, using a Gaussian distribution of mean 4 and standard deviation 1, whereas for $y = 0$ we generate 500 and 2500 cases (for the small and large datasets), respectively, using a Gaussian distribution of mean 2 and standard deviation 1. To test the effects of dataset shift, the mean of the positive class was varied from 4 (almost perfect separation between both training patterns) until 2 (complete overlap) using homogeneous decreases of the mean of the positive cases of 0.25 (ten experiments in total).

Figures 1 and 2 show the results of the AUC (Area Under the ROC Curve) for the new sample, using the proposed strategy and the alternative approaches Stability Index and Kolmogorov Smirnov Test. The X-axis of each plot presents the mean of the positive cases of the new sample, while the Y-axis shows the bounds of the DCI_β (the dotted lines for $z_\alpha = 1$, the dashed lines for $z_\alpha = 2$, and the dash-dot lines for $z_\alpha = 3$), which is centered in 1 (no shift, $\mu_o = \mu_n = 4$). Each circle in the plot represents the mean ratio, while the vertical lines are the

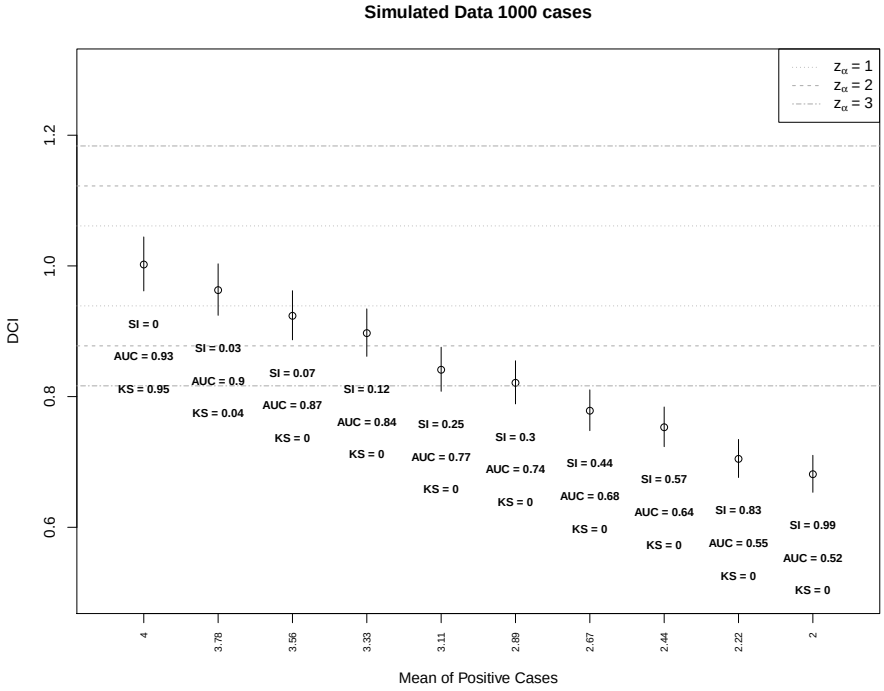


Figure 1: Simulated data with higher uncertainty (1000 cases).

confidence intervals obtained from Fieller's Theorem.

The main difference between both plots is that, in the case of higher uncertainty in the model coefficients (Figure 1), we observe very broad confidence zones for DCI_β, detecting shifts only in extreme cases. In contrast, the confidence zones are much tighter in the case of lower uncertainty (Figure 2). The reasoning behind this is that if we have a higher uncertainty over the original coefficients that created the model, then we need a stronger evidence of dataset shift in order to suggest a recalibration of the model. When there is a greater certainty over the "true" parameter value, then less evidence is necessary to detect with confidence that concept drift has occurred.

As displayed in Figure 1, a slight shift is detected with our approach for a mean of 3.33 (fourth experiment), where the AUC has dropped to 0.83. In this case, the Stability Index suggests a very slight shift (SI=0.12) and KS already rejected the null hypothesis, suggesting dataset shift from the second experiment onward (p-value=0). A relevant change is identified with our approach for a

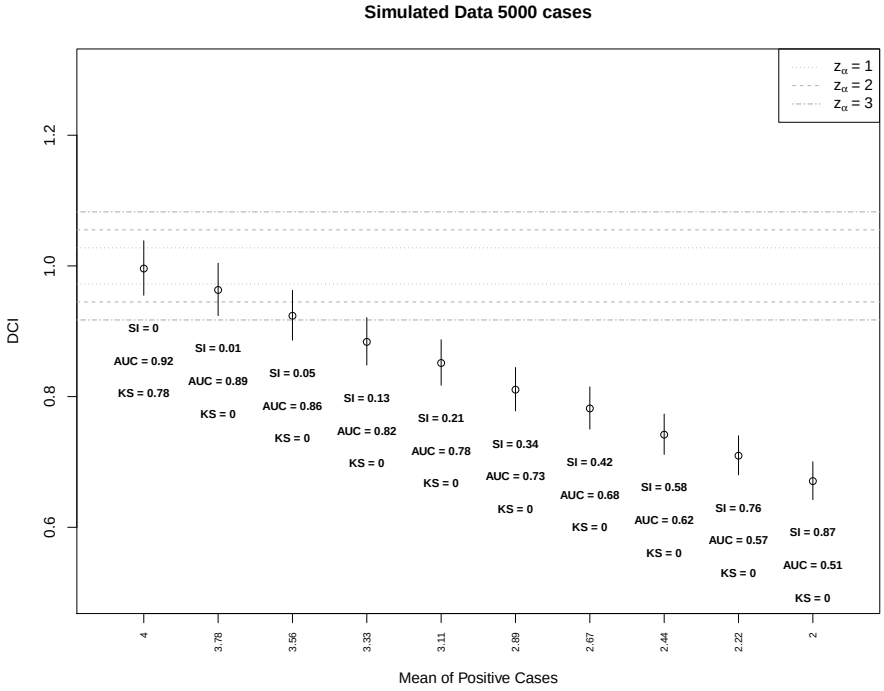


Figure 2: Simulated data with low uncertainty (5000 cases).

mean of 2.89 (sixth experiment), with an AUC of 0.74, which is also consistent with the Stability Index (SI=0.34), since both methods suggest the recalibration of the model. A severe change is detected with our approach for a mean of 2.67 (seventh experiment), which leads to an AUC of 0.68, an unacceptable loss in predictive power, close to 30%, which is also detected by the Stability Index (SI=0.51). For the case of less uncertainty, given by a bigger sample (Figure 2), the confidence zones for our approach are tighter and therefore changes are detected earlier in our experiments (in the third, fourth and fifth experiment for z_α=1, 2, and 3 respectively), while the alternative tests remains with similar conclusions since they are independent from the size of the samples.

Compared with the Stability Index, both tests agree in the occurrence of drastic changes in most cases. However, our approach detects an important dataset shift earlier than SI for the second case (5000 instances). While our approach recommends a recalibration by the fourth experiment, only a minor change is identified by the SI at this point. Considering that the AUC drops

Table 1: Number of cases and defaulters per time cut

	Non-Defaulters	Defaulters	Total
2000 - 2004	21,157	13,818	5383
2005 - 2007	3573	1810	34,975
Total	24,730	15,628	40,358

from 0.92 to 0.83 for a mean of 3.3 in the positive cases, we consider that these changes should be detected, as our method does. By contrast, the KS Test rejects the hypothesis of equal distributions with a p-value close to zero in all cases, detecting dataset shift in cases with only very slight changes, which is not useful when deciding the moment to recalibrate.

Experiments on Real Data

We applied the proposed dataset shift strategy to a real-world credit scoring problem. The dataset contains 40,000 personal loans granted by a financial institution during the period 2000-2007, with 38% of the borrowers being defaulters. The very high percentage of defaulters is explained due to the fact that the government institution granting loans was very lenient in their granting practices, with a 1% rejection rate. A detailed description of this dataset can be found in Bravo *et al.* (2013). We use the loans granted during the period 2000-2004 for constructing a predictive model with the highest possible predicting capabilities using logistic regression. We contrasted this dataset (reference sample) with the instances from 2005 to 2007 (new sample). The imbalance and composition of both samples can be seen in Table 1.

The model was trained using variables that arise from the socio-economic condition of the borrower, and the information that can be collected from the loan application. Variables include the age of the borrower, whether there are guarantors or collateral available, and the economic sector of the borrower, among others. For an in-depth discussion of the variables that appear on a model trained in a similar dataset, and the treatment these variables received, please see Bravo *et al.* (2013). For comparison purposes, a second model was constructed using the new sample. The coefficients of both models are presented in Table 2a and Table 2b.

From Table 2a and Table 2b we observe an evident dataset shift. Several

Variable	Estimate	Std.
Var1	-0.35	0.03 ***
Var2	0.35	0.02 ***
Var3	-0.38	0.04 ***
Var4	0.16	0.02 ***
Var5	0.41	0.03 ***
Var6	0.10	0.05 *
Var7	-0.20	0.05 ***
Var8	0.18	0.04 ***
Var9	-0.47	0.04 ***

(a) Original model

Variable	Estimate	Std.
Var1	-0.12	0.09
Var2	0.44	0.05 ***
Var3	-0.31	0.13 **
Var4	0.07	0.06
Var5	0.54	0.10 **
Var6	0.21	0.19
Var7	0.46	0.18
Var8	-0.03	0.09
Var9	-0.11	0.10

(b) Model trained with new data.

coefficients have either become non-significant or they have changed in magnitude. Notice that the labels for the new sample are normally not available at this stage, so the new model would not be available in a real-life setting, but it provides an interesting experiment when an undetected shift may lead to a significant drop in performance.

Now we study the behavior of the proposed test for the variables in this dataset. The results of three representative cases are presented in Figures 3 to 5 (Variables 1 to 3), while the six remaining variables (Variables 4 to 9) are presented in Appendix .

For variables 1 and 4 (the first one is shown in Figure 3), our approach detects a relevant shift, since the FSM completely leaves the zone of maximum allowed deviance (lower dashed line) for some of the periods of study. For variable 9, a severe shift is detected. For all these variables, their coefficients in Table 2b are non-significant, and the estimated value differs strongly from the original estimated parameter, so that the original parameter is outside the 95% interval of confidence of the new one. As we observe in Figure 3, SI fails to identify this shift in most periods ($SI < 0.25$), while KS adequately detects significant changes (p-value close to zero), but overreacts to small ones.

When a slight shift is detected (Variables 3, 5, 6 and 7, Variable 3 in Figure 4), it can be associated either with a significant parameter that is outside, or almost outside the original confidence intervalzone, or with a non-significant parameter. For variable 3, for example, SI detects major changes in the first period ($SI = 0.44$), and KS consistently rejects the null hypothesis of equal distributions. According to Table 2b, however, this shift does not affect the coefficient of the new model and, as can be observed from Figure 4, this shift is not a major one.

Finally, when no shift is detected (Variables 2 and 8; Variable 2 shown in

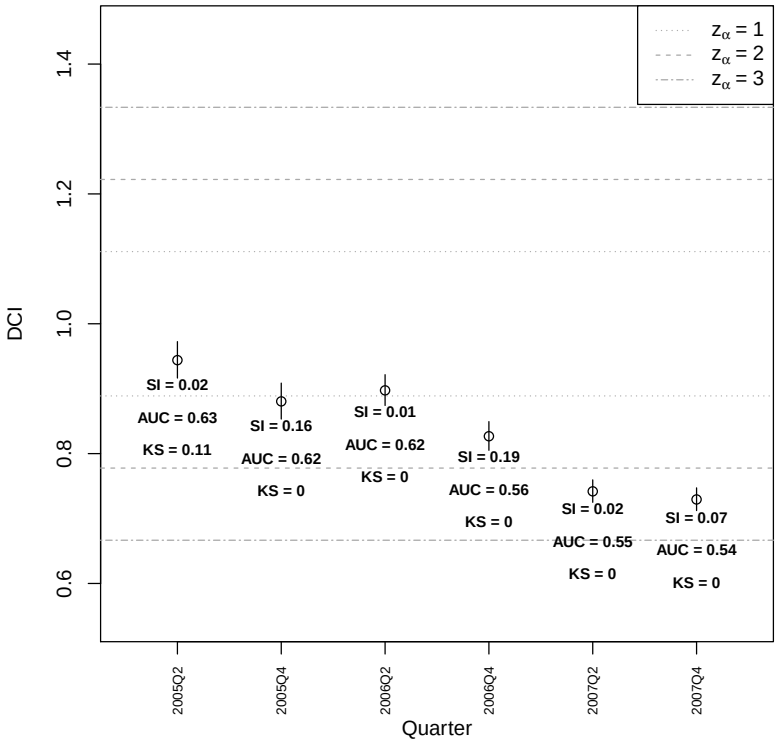


Figure 3: Large shift (Var1).

Figure 5) the parameter has not had a significant shift at all. In all these cases the new estimated parameter is within the confidence zone of the original parameter. As we observe in Figure 5, SI is consistent with the proposed test, while KS rejects the null hypothesis of equal distributions in most cases. This result proves the ineffectiveness of KS, since it sets off the alarms for dataset shift even with minor changes in the means.

For all relevant and severe shift cases, the coefficients in Table 2b are non-significant, and their estimated values differ significantly compared with the estimations made with the reference sample. When a slight shift is detected, we also observe important changes in two out of three variables. Finally, the coefficients associated with variables where no shift is found remain relatively stable in the reference sample.

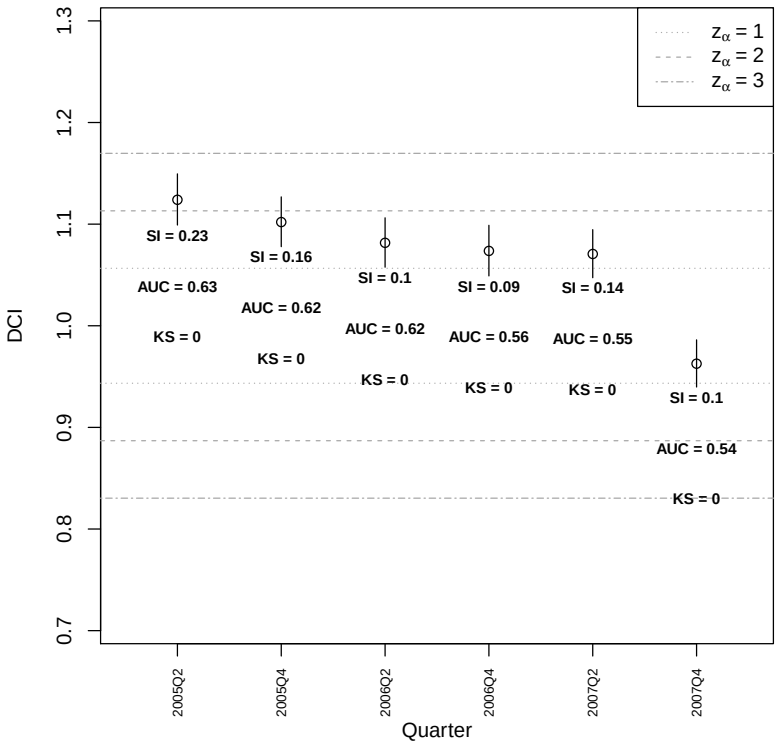


Figure 4: Slight shift (Var3).

We conclude from previous experiments that major changes in the dataset can be identified successfully with our proposal, while small changes are not labeled as dataset shift, in contrast to approaches such as KS. It is also important to notice that the confidence zone for the maximum deviance allowed (the horizontal lines in the plots) does not depend on only the sample size, and for variables with more uncertainty in their coefficients, a significant shift should occur to set off an alarm.

FSM differs from SI when the model itself presents instability. Sometimes the change is not big enough to ascertain that drift has occurred, and so FSM does not set off any alarms, while SI shows contradictory results. Another example is when our approach detects a relevant shift, while SI only sets off a slight change alarm, which occurs in all cases when the parameter associated with the

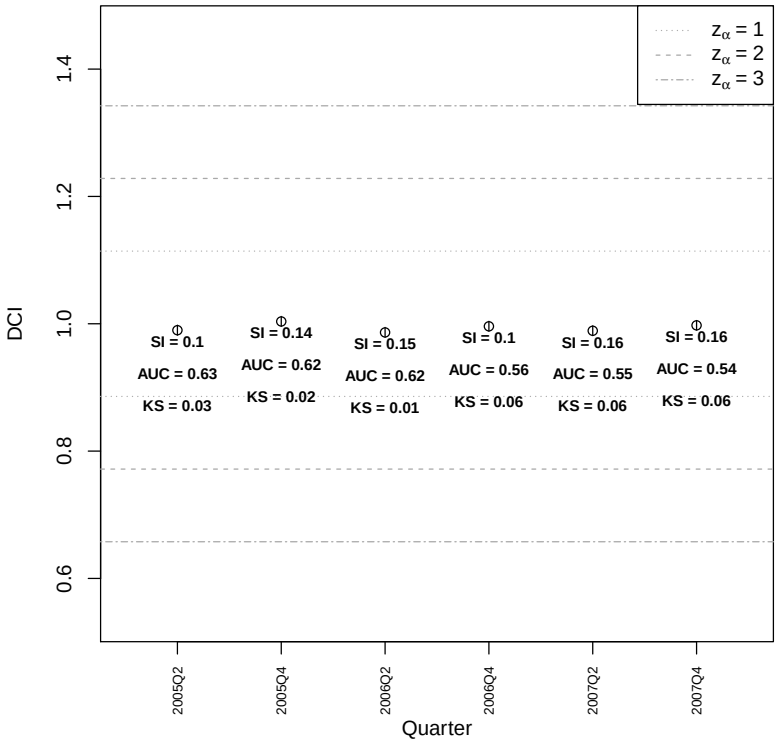


Figure 5: No shift (Var2).

variable is uncertain, so more tolerance is allowed for the variable. However, whenever either SI or FSM show that no shift has occurred, or that a large shift has occurred, then the other measure tends to agree. Additionally, it can be seen that a deterioration of the AUC measure occurs as time progresses, which is to be expected, and some of that deterioration can be assigned to concept drift.

Conclusions

In this paper we have developed a backtesting measure for predictive models that incorporates a maximum allowed shift, constructed from the confidence interval of the original model parameter, for each variable. The Fieller Stability Measure

(FSM) is estimated using the Fieller confidence interval, a test frequently used in medicine for cost comparison, that we show is also appropriate for concept shift.

Similar to the Stability Index, which is widely used in business applications, the test gives three different severities for concept shift: a slight, a relevant, and a severe shift alert when there is an empty intersection of the FSM and the maximum allowed deviation constructed from one, two, and three standard deviations over the mean of the original parameters, respectively. The main difference with previous measures is that the FSM measure calculates the limits from the parameters of the original model, and thus reflects the trade-off between more certain estimation (given by a large dataset, for example) and the robustness of the model when facing concept shift. This procedure leads to two important advantages: first, the bounds for the alerts are grounded in statistical measures and are not set arbitrarily; and second, the model takes additional elements into account, such as certainty in the original coefficients, sample sizes, and sample distributions (given by the first two moments of the original and new datasets).

When the procedure was applied to simulated and real-world datasets, the positive performance and robustness of our approach became apparent. Compared with SI, both tests agree in most cases. Our approach, however, present more stable and consistent results, while SI can be affected by the binning process and deliver inaccurate values. If there are discrepancies, our approach performs better in predicting which variables will change their coefficients drastically in a recalibration of the model. Compared with traditional statistical tests, such as KS, our approach consistently ignores minor changes in the distribution, while KS tends to identify dataset shift even with marginal variations which are normal in business applications and should not affect the performance of the model.

As a final conclusion, the proposed measure is very simple to estimate, but at the same time it is very powerful because of the information that it provides to the modeler. Since it can be shown readily on a graph, it is also very simple to use; furthermore, it can be applied to all models which have asymptotically normal parameters, and for variables with any type of distribution. As such, it can be proposed as a business tool for evaluating concept shift that is easily deployed.

Acknowledgements

The first author acknowledges the support of CONICYT Becas Chile PD-74140041. The second author was supported by CONICYT FONDECYT Initiation into Research 11121196. Both authors acknowledge the support of the Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16).

References

Anderson, R. *The Credit Scoring Toolkit*. Oxford University Press, 2007.

Baesens, B. *Analytics in a Big Data World*. John Wiley and Sons, New York, USA., 2014.

Baesens, B., Mues, C., Martens, D. and Vanthienen, J. 50 years of data mining and or: upcoming trends and challenges. *Journal of the Operational Research Society* **60**(S1):16–23, 2009.

Basu, A., Harris, I. R. and Basu, S. *Handbook of Statistics*, volume 15: Robust Inference, chapter Minimum distance estimation: The approach using density-based distances., pages 21–48. Elsevier, 1997.

Bergtold, J., Yeager, E. and Featherstone, A. Sample size and robustness of inferences from logistic regression in the presence of nonlinearity and multicollinearity. In: *Proceedings of the Agricultural & Applied Economics Associations 2011 AAEA & NAREA Joint Annual Meeting*. Pittsburg, Pensilvania, USA, 2011.

Beyene, J. and Moineddin, R. Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Medical Research Methodology* **5**(32):1–7, 2005.

Birón, M. and Bravo, C. *Data Analysis, Machine Learning and Knowledge Discovery*, chapter On the Discriminative Power of Credit Scoring Systems Trained on Independent Samples, pages 247 – 254. Springer International Publishing, 2014.

Bravo, C., Maldonado, S. and Weber, R. Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research* **227**(2):358–366, 2013.

Castermans, G., Hamers, B., Van Gestel, T. and Baesens, B. An overview and framework for PD backtesting and benchmarking. *The Journal of the Operational Research Society* **61**(3):359–373, 2010.

Cieslak, D. and Chawla, N. Detecting fractures in classifier performance. In: *Proceedings of the Seventh IEEE International Conference on Data Mining*, pages 123–132. Department of Computer Science and Engineering, University of Notredame, Indiana, USA, 2007.

Fieller, E. C. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* **16**(2):175–185, 1954.

Hofer, V. and Krempl, G. Drift mining in data: A framework for addressing drift in classification. *Computational Statistics and Data Analysis* **57(1)**:377–391, 2013.

Hosmer, D. and Lemeshow, H. *Applied Logistic Regression.* John Wiley & Sons, 2000.

Kelly, M., Hand, D. and Adams, N. The impact of changing populations on classifier performance. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 367–371. San Diego, California, USA, 1999.

Lewis, E. M. *An Introduction to Credit Scoring.* Fair, Isaac & Co., Inc, California, USA., 1992.

Lima, E., Mues, C. and Baesens, B. Monitoring and backtesting churn models. *Expert Systems with Applications* **38**(1):975–982, 2011.

Moreno-Torres, J. G., Raeder, T. R., Aláiz-Rodríguez, R., Chawla, N. V. and Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognition* **45**(1):521–530, 2012.

Quiñonero Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D. (eds.) *Dataset Shift in Machine Learning.* MIT Press, 2009.

Robinson, S., Brooks, R. and Lewis, C. Detecting shifts in the mean of a simulation output process. *Journal of the Operational Research Society* **53**(5):559–573, 2002.

Schenker, N. and Gentleman, J. F. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* **55**(3):182 – 186, 2001.

Schlimmer, J. and Granger, R. Beyond incremental processing: tracking concept drift. In: *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 502–507. San Francisco, CA, USA, 1986.

Siddiqi, N. *Credit risk scorecards: developing and implementing intelligent credit scoring.* John Wiley and Sons, 2006.

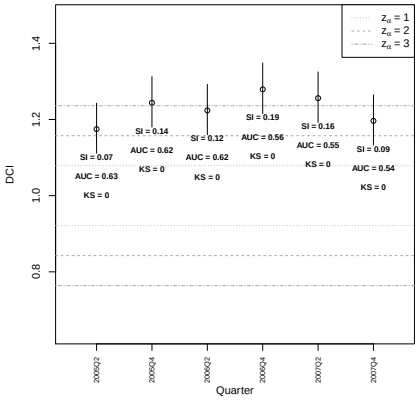
Smirnov, N. Tables for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19**(2):279–281, 1948.

Utts, J. M. and Heckard, R. F. *Mind on Statistics.* Cengage Learning, 4th edition, 2012.

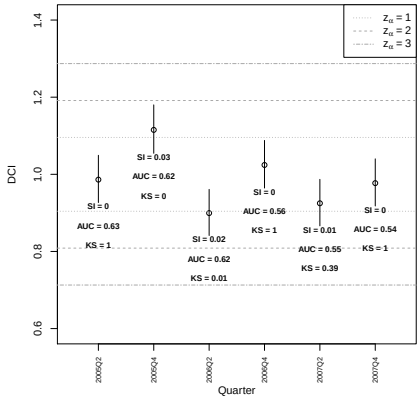
Yang, Y., Wu, X. and Zhu, X. Conceptual equivalence for contrast mining in classification learning. *Data and Knowledge Engineering* **67**(3):413–429, 2008.

Appendix A. Results for the Experiments on Real Data

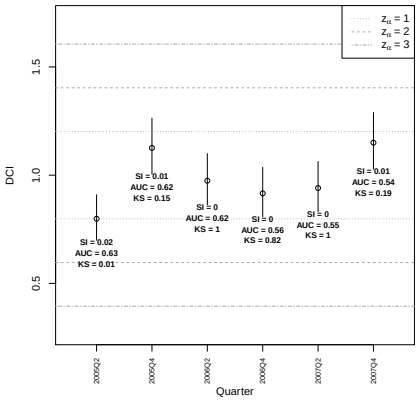
The following graphs present the results of the dataset shift tests (the proposed approach and SI) for the six remaining variables (Variables 4 to 9).



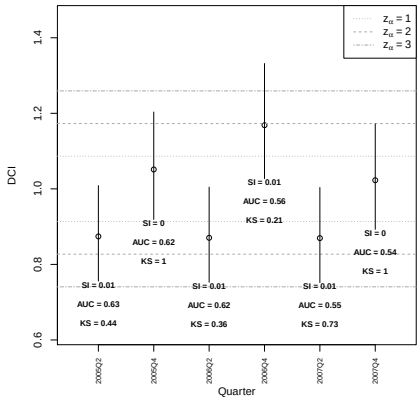
(a) Relevant shift (Var4).



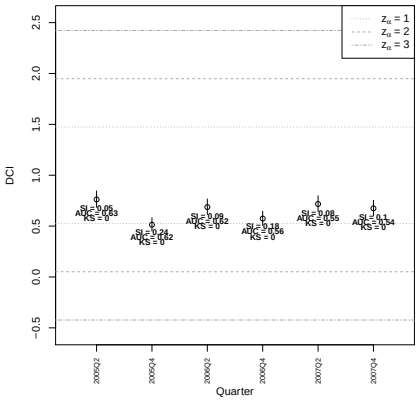
(b) Slight shift (Var5).



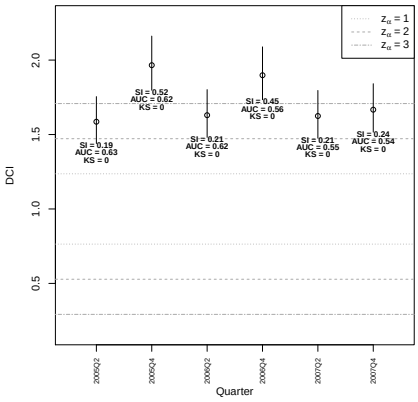
(c) Slight shift (Var6).



(d) Slight shift (Var7).



(e) No shift (Var8).



(f) Severe shift (Var9).