Volume 36 (1), pp. 19–34
http://orion.journals.ac.za

ORiON
ISSN 0259–191X (print)
ISSN 2224–0004 (online)

On testing the hypothesis of population stability for credit risk scorecards

J du Pisanie* IJH Visagie†

Received: 19 May 2020; Accepted: 15 June 2020

Abstract

Scorecards are models used in credit risk modelling. These models segments a population into various so-called "risk buckets" based on the risk characteristics of the individual clients. Once a scorecard has been developed, the credit provider typically prefers to keep this model in use for an extended period. As a result, it is important to test whether or not the model still fits the population. To this end, the hypothesis of population stability is tested; this hypothesis specifies that the current proportions of the population in the various risk buckets are the same as was the case at the point in time at which the scorecard was developed. In practice, this assumption is usually tested using a measure known as the population stability index (which corresponds to the asymmetric Kullback-Leibler discrepancy between discrete distributions) together with a well-known rule of thumb.

This paper considers the statistical motivation for the use of the population stability index. Numerical examples are provided in order to demonstrate the effect of the rule of thumb as well as other critical values. Although previous numerical studies relating to this statistic are available, the sample sizes are not realistic for the South African credit market.

The paper demonstrates that the population stability index has little statistical merit as either a goodness-of-fit statistic to test the hypothesis of population stability or as an intuitive discrepancy measure. As a result, a novel methodology for testing the mentioned hypothesis is proposed. This methodology includes a restatement of the hypothesis to specify a range of "acceptable" deviations from the specified model. An alternative test statistic is also employed as discrepancy measure; this measure has the advantage of having a simple heuristic interpretation in the context of credit risk modelling.

Key words: Goodness-of-fit testing, hypothesis testing, population stability, risk analysis.

*Monocle Solutions, South Africa, email: johan.dupisanie@monocle.co.za
†Corresponding author: School of Mathematical and Statistical Sciences, North-West University, South Africa, email: jaco.visagie@nwu.ac.za

http://dx.doi.org/10.5784/36-1-678

1 Introduction and motivation

Credit risk scorecards are used extensively in banks as well as many other institutions, including retailers, micro lenders and collections agencies. The principal aim of scorecards is to segment the population into a number of classes (often referred to as risk buckets) based on the risk characteristics of the individuals making up the population. Consider, for example, the case where the population is divided into ten classes, each containing 10% of the population. The first class contains the 10% of the population with the lowest probability of default (or, alternatively, the highest rate of response in the case of marketing or collections scorecards), and so on. Scorecards are mainly used in three areas of banking; credit risk scoring, collections scoring and marketing scorecards. Each of these are briefly considered below.

Credit risk scorecards provide a robust method of scoring and segmenting a population for risk purposes. The risk segment that a customer falls into is often directly used in the approval of credit as well as the setting of a credit limit, the interest rate and even the products offered to the customer. Scorecards are popular in credit risk modelling due to their simple interpretation, which lends itself to explanation to a non-technical audience. This is a substantial advantage over models utilising, for example, neural networks. The inherent repeatability associated with scorecards also ensures that a customer will be approved (or declined) consistently, unless their underlying risk profile changes.

Collections scorecards provide collections teams with a stable tool to determine which customers are likely to make repayments. Through segmentation of the population into different classes, strong collections strategies can be built which may include contacting only those customers for whom the probability of successful collections exceeds a specified threshold, ensuring that the costs associated with collections are only incurred when offset by the likelihood of repayment.

The final major use of scorecards is found in marketing. In the South African context, scorecards are often used to segment the population in order to determine which segments will be most likely to take up a product. This information can be used to determine which marketing strategy will likely lead to the best results. The South African credit industry actively tracks all credit customers, providing vast amounts of data that can be used for the building of marketing scorecards.

Building a scorecard is often expensive and time consuming. As a result, financial institutions prefer to keep a scorecard in use for an extended period, given that the scorecard provides a sufficiently accurate representation of the population throughout the time that it is in use. The current paper is concerned with testing the hypothesis that the population considered has remained stable over time, meaning that a scorecard which was developed at an earlier date is a sufficiently accurate model of the population at present. In practice, a measure referred to as the population stability index (PSI) is often used in order to make this determination.

In order to determine whether or not a given scorecard remains a representative model, the present composition of the population is compared to the composition of the population at the time when the scorecard was developed. The PSI constitutes a discrepancy measure between these two compositions. The hypothesis of population stability and the PSI are

of considerable practical interest and is available in the statistical software packages SAS and R, see [6] and [1] respectively. An investigation into the statistical properties of the PSI can be found in [10], while details regarding the practical application of the PSI are available in [8] and [9].

The aim of this paper is to provide practical insight into the testing of the hypothesis of population stability. As a result, the technical details are kept to a minimum and heuristic explanations and examples are favoured over a rigorous mathematical exposition.

Testing the hypothesis of population stability constitutes a goodness-of-fit test. For more details regarding goodness-of-fit testing; see, for example, [5]. In this paper, several properties of the PSI which are not typically associated with a goodness-of-fit test statistic are illustrated, and arguments are presented in favour of the use of a different test statistic. Furthermore, an alternative formulation of the hypothesis of population stability is provided which the authors believe is of greater practical relevance.

It should be noted that the PSI is typically not the only measure employed by practitioners when making a determination regarding the continued validity of a scorecard. Another consideration that may be of importance is that of the practical significance of the observed change in the distribution, as measured by an effect size. The large sample sizes often available in practice may result in statistically significant differences between two populations even in the absence of practically significant differences. However, while other measures are also taken into account, the value of the PSI plays an important role in determining whether or not a scorecard can remain in use.

The remainder of the paper is structures as follows. Section 2 details the hypothesis of population stability. Special attention is paid to the PSI in this section, including a discussion of the properties of the PSI as an intuitive discrepancy measure. Section 3 contains the newly proposed methodology for testing the hypothesis of population stability. This section proposes a reformulation of the hypothesis being studied as well as an alternative test statistic. Some conclusions are presented in Section 4.

2 The hypothesis of population stability

Scorecards can become unusable for a number of reasons. These include, but are not limited to, the following:

1. The current composition of the population no longer resembles that of the population that the scorecard was built on. This can happen due to underlying changes in the population, for instance more younger people applying for the product in question; a larger proportion of affluent customers applying for the product, *etc.*
2. The population remains stable, but the inherent risk in the applicants change. This can happen when the economy goes through an upswing or downswing when compared to the development period of the scorecard.
3. The product offering changes and thereby attracts a new type of client. This could force the population characteristics to veer away from the original characteristics that were identified when building the scorecard.

The event described in 2. would not result in a change in the proportion of the population in each of the risk buckets. However, the inherent risk associated with an individual with a fixed credit score would change. It is important to note that, even though the scorecard can no longer be used in the same way as it was before the change, the population has indeed remained stable over time. As a result, this situation does not necessarily imply a violation of the assumption of population stability. Rather, other measures (specific to the business at hand) will have to be used in order to detect a change of this type. Consequently, the current paper is not concerned with detecting changes such as those specified in event 2. and assumes throughout that the levels of risk associated with a fixed credit score remains unchanged.

When the events listed in 1. and 3. above occur, the proportion of the population associated with each risk bucket will deviate from that of the target proportion (specified by the scorecard). A discrepancy measure of some kind (such as the PSI) is required in order to measure the magnitude of the deviation between the proportions specified by the model and the observed proportions. Based on this measure, a determination as to whether or not the observed deviation is within tolerable limits is required to be made. If the deviation is not within tolerable limits, then the assumption of population stability is rejected and a new scorecard has to be developed (or some adjustment has to be made in order to justify the use of the current scorecard in some way).

The hypothesis of population stability to be tested is that the current proportion of clients in each risk bucket is correctly specified by the model. Let the model consist of k risk buckets and denote the population proportions associated with these buckets by

$$\mathbf{q} = (q_1, \ldots, q_k).$$

Let

$$\mathbf{p} = (p_1, \ldots, p_k) \quad \text{and} \quad \mathbf{P} = (P_1, \ldots, P_k),$$

respectively, denote the population and observed (random) sample proportions associated with the different risk buckets. Note that $\mathbf{q}$ is determined when the scorecard is developed while $\mathbf{P}$ is obtained when the hypothesis of population stability is tested. The hypothesis of population stability can be formally expressed as follows:

$$H_0 : p_j = q_j, \forall j \in \{1, 2, .., k\} \quad \text{versus} \quad H_A : p_j \neq q_j, \text{ for some } j \in \{1, 2, .., k\}. \tag{1}$$

This hypothesis is to be tested using the observed sample quantities in $\mathbf{P}$. The PSI is a popular discrepancy measure used in order to test the hypothesis specified in (1). This measure is considered in more detail below.

2.1 Calculation of the PSI

Using the notation defined above, the PSI can be defined as follows:

$$\Psi(\mathbf{q}, \mathbf{P}) = \sum_{j=1}^{k} (P_j - q_j) \log\left(\frac{P_j}{q_j}\right). \tag{2}$$

Note that Ψ is a discrepancy measure between the observed proportion of observations in each bucket and the corresponding expected proportions under the model. The hypothesis

of population stability is rejected for large values of the test statistic in (2). A simple rule of thumb, suggested in [8], is widely used in practice. If $\Psi \in [0, 0.1)$, then the population is deemed to have remained stable. If $\Psi \in [0.1, 0.25)$, then a small shift in the population has occurred and further investigation is required in order to ascertain whether or not the model is still accurate. If $\Psi \geq 0.25$, then a large change in the characteristics of the population has occurred and the model is no longer valid.

The PSI corresponds to the well-known Kullback-Liebler divergence defined in [4]. Note that several versions of this information criterion can be found in the relevant literature; the PSI corresponds to the asymmetric measure of the difference between two discrete probability distributions. The central role of the Kullback-Liebler divergence measure in information theory is well-known and the interested reader is referred to [3] for more details.

Below, several shortcomings of the PSI as an intuitive measure of difference between two discrete probability distributions are discussed. In each case, the claims made are supplemented with specific numerical examples in order to give credence to these assertions. The following numerical example is used repeatedly below. Consider the case where a population is segmented into 10 risk buckets of equal size at the time that the scorecard is constructed; *i.e.*

$$\mathbf{q} = (q_1, \ldots, q_{10}) = (0.1, \ldots, 0.1). \tag{3}$$

This configuration is often found in practice; see [10]. Analyses relating to this configuration are provided below.

2.2 The effect of the sample size

In (3), the expected proportion of clients in each risk bucket is 10%. Consider the case where the observed population is classified into risk buckets and the proportion of clients in one bucket is 5% and the proportion in another is 15%, while all remaining buckets contain 10% of the population. Denote this configuration by

$$\mathbf{b_1} = (\cdot, \cdot, \cdot, \cdot, 5\%, 15\%, \cdot, \cdot, \cdot, \cdot),$$

where $\cdot$ is used as shorthand notation for 10%.

The aim of the PSI is to ascertain whether or not the observed deviation between the observed and expected proportions can be ascribed to statistical variation. The PSI aims to reject the assumption of population stability in the case where the observed proportions differ from the proportions specified by the model more drastically than can be ascribed to random chance. In the example currently under consideration, the PSI is calculated to be

$$\Psi(\mathbf{q}, \mathbf{b_1}) = 0.055,$$

which does not exceed 0.1 or 0.25, and, as a result, the hypothesis of population stability is not rejected when using the rule of thumb, regardless of the sample size. However, deviations of this magnitude may well be commonplace under the null hypothesis in the case of a relatively small sample while being unrealistically large for large samples. This assertion is discussed in more detail below.

The critical points 0.1 and 0.25 do not possess any inherent statistical interpretation. When performing a goodness-of-fit test, such as the test specified in (1), the rejection region of the test statistic is typically chosen with reference to one or more quantiles of the distribution of the test statistic used under the null hypothesis. This ensures that the probability of a Type I error (the probability of falsely rejecting the null hypothesis) is known *a priori*. In order to demonstrate that the test has merit, it is usual to examine the power of the test against a range of alternatives, often via Monte Carlo simulation.

The fact that the critical points used in conjunction with the PSI are not chosen with reference to the quantiles of the null distribution of this statistic, means that the modeller does not have control over the probability of a Type I error. Furthermore, the quantiles of the null distribution of a goodness-of-fit test statistic are typically functions of the sample size (denoted by n); this is the case for the PSI. As a result, the Type I error probabilities associated with a cut-off point of 0.1 or 0.25 are functions of the sample size. This dependence was illustrated in [10]. However, the maximum sample size considered was 1600, which is an unrealistically small sample when considering the South African credit market (as well as many other markets around the world). In this paper, these (and other) probabilities are considered for larger sample sizes.

For a given sample size, the probability of a Type I error associated with the proposed cut-off points of 0.1 and 0.25, can be approximated using Monte Carlo simulation. Under the null hypothesis, for a sample of size n, the joint distribution of the number of items in each risk bucket is $multinomial(n, \mathbf{q})$. As a result, the algorithm below can be used to estimate the probability of a Type I error for a given sample size and cut-off point.

The Type I error probability can be estimated using the following algorithm:

Algorithm 1: Algorithm for estimating the probability of a Type I error

1. Generate the number of items in each risk bucket from a $multinomial(n, \mathbf{q})$ distribution. Denote the resulting vector by $\mathbf{X}$.
2. Given the value of $\mathbf{X}$, calculate the proportion of items in each risk bucket by dividing each element of $\mathbf{X}$ by the sample size n; $\mathbf{P} = \mathbf{X}/n$.
3. Given the value of $\mathbf{P}$, calculate the observed PSI value; $\Psi(\mathbf{q}, \mathbf{P})$.
4. Repeat steps (1) to (3) MC times in order to obtain MC realisations of the PSI under the null hypothesis. Denote the resulting values by $\Psi_1, \ldots, \Psi_{MC}$.
5. Determine the probability of a Type I error, denoted by α, (using a cut-off point of c) as follows:

$$\alpha = \frac{1}{MC} \sum_{j=1}^{MC} \mathrm{I}(\Psi_j \geq c),$$

where $\mathrm{I}(\cdot)$ denotes the indicator function.

In the South African credit market, sample sizes ranging from 10 000 to 100 000 are frequently observed. Algorithm 1 was used to approximate the probability of a Type I error for sample sizes $n \in \{10^3, 10^4, 10^5, 10^6\}$ using 1 million Monte Carlo replications in each case. The results obtained are identical; not a single exceedance of either of the critical

points for any of the sample sizes considered. As a result, the probability of exceeding either threshold is estimated to be 0 for each sample size considered. This means that, for practical purposes, it is impossible to falsely reject the hypothesis in (1) using the rules of thumb in [8] in the case of realistic sample sizes for the South African credit market.

As was mentioned above, the critical region of a goodness-of-fit test is typically chosen with reference to a quantile of the null distribution of the test statistic given the sample size. Table 4, below, shows the estimated (rescaled) critical values for the PSI for the various sample sizes mentioned; the entries in the table are the relevant quantiles multiplied by the sample size. In the tables below, Q_γ denotes the quantile of the null distribution such that

$$F(Q_\gamma) = \gamma,$$

where F denotes the distribution function of Ψ. Again, the reported estimates are obtained using 1 million Monte Carlo replications. Typically, the hypotheses of this kind are tested at the 1%, 5% or 10% levels of significance, the table shows the critical values for each of these levels. All numerical calculations were performed in the software package R; see [7]. Tables were formatted using the "stargazer"; see [2].

n	$nQ_{0.90}$	$nQ_{0.95}$	$nQ_{0.99}$
10^3	14.79	17.05	21.90
10^4	14.68	16.91	21.67
10^5	14.68	16.92	21.63
10^6	14.70	16.92	21.63

Table 1: *Scaled critical values of the PSI for various sample sizes.*

While the stability of the rescaled critical values is remarkable, it is not unexpected. The numerical results suggest that $n\Psi$ converges to a non-degenerate limit distribution and, due to the large sample sizes considered, the null distribution of $n\Psi$ can be accurately approximated by its limit null distribution.

In order to obtain the critical values (corresponding to the various nominal rates of significance) from the tables, one needs to divide the relevant table entry by the sample size in the leftmost column. This provides strong evidence to the claim that the sample size influences the quantiles of the null distribution. In order to further illustrate the effect of the sample size on the null distribution, consider Figures 3 and 4. Figure 3 shows the critical value for various sample sizes ranging from 10^3 to 10^5 (note that the x-axis is shown in logarithmic scale), while Figure 4 illustrates kernel density estimates of the null distributions of the PSI for sample sizes $n = 10^4$ (dashed line) and $n = 10^5$ (solid line), respectively.

2.3 On the effect of the number of risk buckets

As was the case for the sample size, the number of risk buckets used in the model also has a pronounced effect on the null distribution of the PSI. Consider again the model specified in (3). The model does not specify the exact distribution of the risk scores, rather it specifies the proportion of scores within each bucket. As a result, an increase in

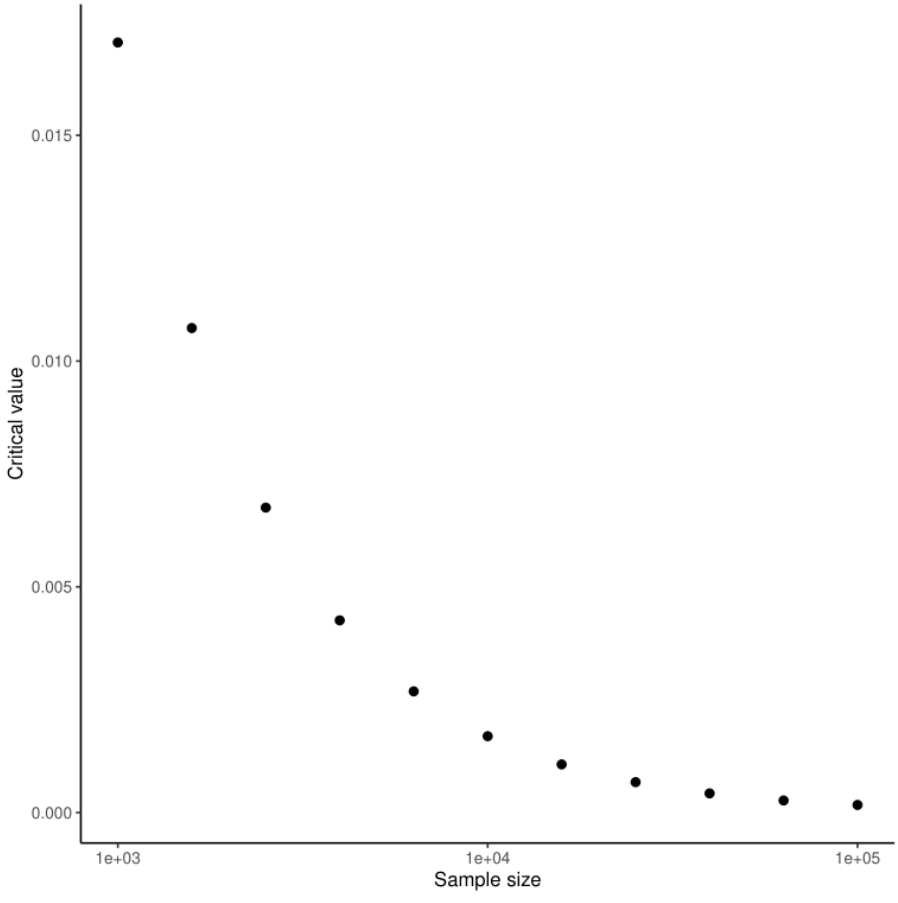

Figure 1: *Critical values as a function of sample size.*

the number of buckets used corresponds to greater specificity in the model. For example, if a shift in the risk scores take place such that the current proportions are as follows:

$$\mathbf{b_2} = (5\%, 15\%, 5\%, 15\%, 5\%, 15\%, 5\%, 15\%, 5\%, 15\%),$$

then the PSI is calculated to be 0.275 and population stability is rejected using the rule of thumb. On the other hand, consider the case where the model consists of only five risk buckets, each obtained by combining two adjacent risk buckets. In the current example, each of the five risk buckets would contain 20% of the population, which corresponds exactly to the specifications of the model. This situation results in an observed PSI value of 0, providing no evidence against the hypothesis of population stability. This situation is considered in more detail below

It was mentioned above that it is commonplace to use 10 risk buckets when modelling credit risk; see [10]. In this example, the sample size is fixed at 10^5. Consider three distinct cases where the population is divided into five, ten and twenty risk buckets, respectively. In each case, the proportions of the population in the risk buckets are equal under the null hypothesis. As before, the Type I error for $k \in \{5, 10, 20\}$ is estimated using 1 million Monte Carlo replications in each case. As before, the suggested cut-off

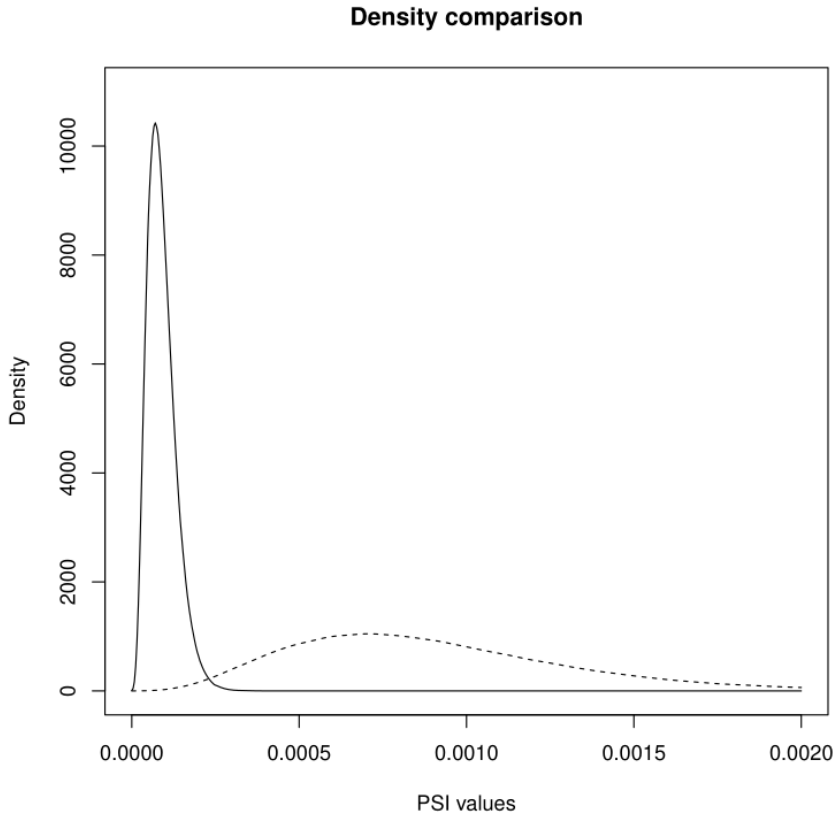

Figure 2: *Kernel density estimates of the null distribution of the PSI for sample sizes $n = 10^4$ (dashed line) and $n = 10^5$ (solid line), respectively.*

points, 0.1 and 0.25, are not exceeded in a single case considered. As was the case when the sample sizes were varied, the probability of exceeding either threshold is estimated to be 0 in each case when using the mentioned rules of thumbs. Again, the critical values associated with nominal significance levels of 1%, 5% and 10% were calculated. Table 2 shows these estimated critical values multiplied by the sample size.

k	$nQ_{0.90}$	$nQ_{0.95}$	$nQ_{0.99}$
5	7.78	9.50	13.34
10	14.68	16.90	21.60
20	27.20	30.16	36.16

Table 2: *Scaled critical values of the PSI for various numbers of risk buckets.*

Figure 3 shows kernel density estimates of the null distribution of the PSI for numbers of risk buckets $k = 10$ (dashed line) and $k = 20$ (solid line), respectively, for a fixed sample size of $n = 10^5$.

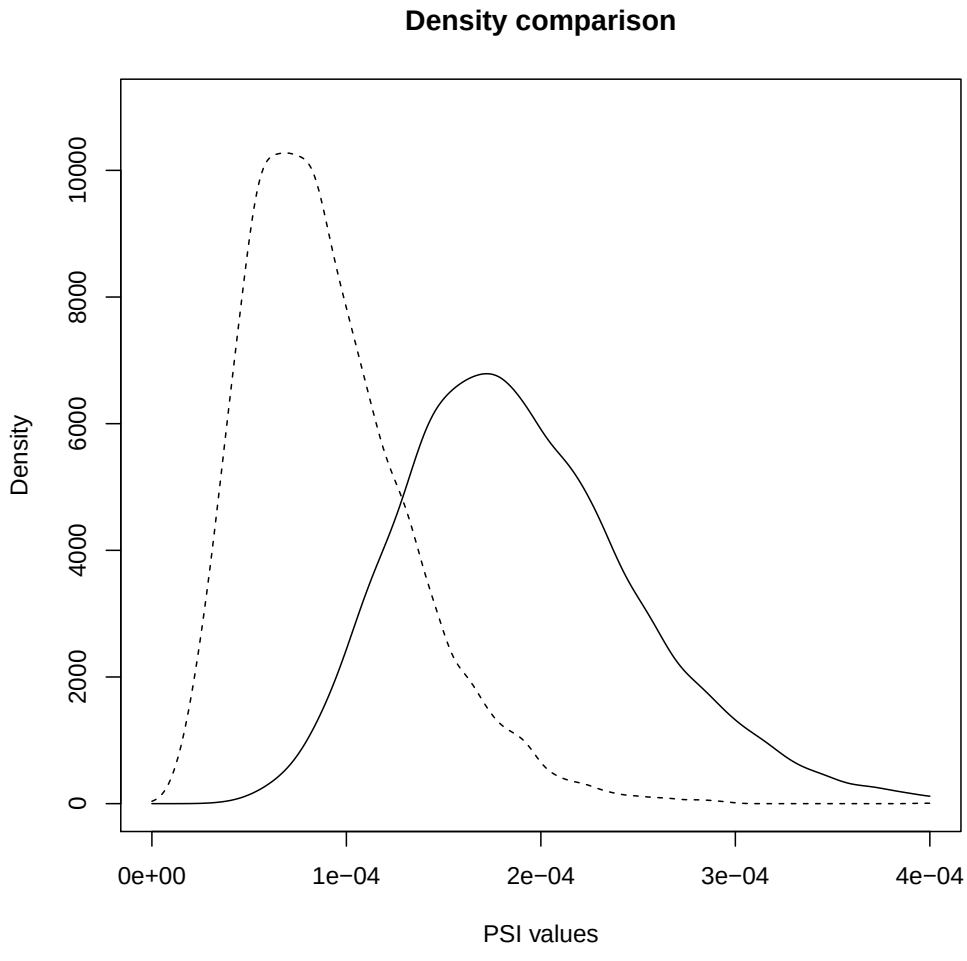

Figure 3: *Kernel density estimates of the null distribution of the PSI for number of risk buckets* $k = 10$ *(dashed line) and* $k = 20$ *(solid line), respectively.*

2.4 On the power of the test

The authors believe that the arguments and examples presented above provide motivation for not using the PSI together with the rule of thumb suggested in [8]. However, up to this point, no argument has been presented against the use of the critical values shown in Tables 1 and 2. On the contrary, these critical values take both the sample size and the number of risk buckets used into account, and these critical values are chosen in such that the probability of a Type I error is known *a priori*. However, the numerical examples below present arguments against the use of the PSI together with these critical values based on the power of the test.

The numerical examples shown above were concerned with investigating the behaviour of the PSI under the null hypothesis. Another important consideration is the probability that the PSI will be able to detect a given deviation from the null hypothesis. To this

end, consider three possibilities for the population at present:

$$\mathbf{b_3} = (0.099, 0.101, 0.099, 0.101, 0.099, 0.101, 0.099, 0.101, 0.099, 0.101), \quad (4)$$

$$\mathbf{b_4} = (0.1, 0.1, 0.1, 0.095, 0.105, 0.1, 0.1, 0.1, 0.1, 0.1), \quad (5)$$

$$\mathbf{b_5} = (0.1, 0.097, 0.103, 0.1, 0.097, 0.103, 0.1, 0.097, 0.103, 0.1). \quad (6)$$

The power of the PSI against the alternatives given by $\mathbf{b_3}$, $\mathbf{b_4}$ and $\mathbf{b_5}$ can be computed using an adaptation of Algorithm 1. For the sake of completeness, this algorithm is provided below.

The power of a test against a distribution with proportions $\mathbf{b}$ can be estimated as follows:

Algorithm 2: Algorithm for estimating the power of the test.

1. Generate the number of items in each risk bucket from a $multinomial(n, \mathbf{b})$ distribution. Denote the resulting vector by $\mathbf{X}$.
2. Given the value of $\mathbf{X}$, calculate the proportion of items in each risk bucket by dividing each element of $\mathbf{X}$ by the sample size n; $\mathbf{P} = \mathbf{X}/n$.
3. Given the value of $\mathbf{P}$, calculate the observed PSI value; $\Psi(\mathbf{q}, \mathbf{P})$.
4. Repeat steps (1) to (3) MC times in order to obtain MC realisations under the null hypothesis. Denote the resulting values by $\Psi_1, \ldots, \Psi_{MC}$.
5. Calculate the power of the test (using a cut-off point of c) as follows:

$$\frac{1}{MC} \sum_{j=1}^{MC} \mathrm{I}(\Psi_j \geq c).$$

Table 3 shows the estimated power against each of the examples at the 10%, 5% and 1% levels obtained using Algorithm 2 (these probabilities are denoted by $P_{0.10}$, $P_{0.05}$ and $P_{0.01}$, respectively). Examples 1, 2 and 3 in the table correspond to the proportions given in (4), (5) and (6), respectively. The sample size is fixed at 10^5.

	$\mathbf{P_{0.10}}$	$\mathbf{P_{0.05}}$	$\mathbf{P_{0.01}}$
Example 1	100.00%	99.99%	99.90%
Example 2	100.00%	99.99%	99.91%
Example 3	100.00%	100.00%	99.96%

Table 3: *Powers of the PSI against the three examples considered.*

The authors believe that the majority of practitioners will agree that none of the examples presented above constitute a substantial change in the population. As a result, the authors believe that most practitioners will be of the opinion that the population has remained stable. However, Table 2 shows that, in all three examples considered, the probability of rejecting the hypothesis of population stability is approximately 100% at each level of significance considered.

The high power of the test against any deviation from the null distribution is due to the large sample sizes considered. This, in conjunction with the arguments presented above, prompts the authors to propose an alternative methodology for testing the hypothesis of population stability. This methodology is discussed below.

3 Proposed methodology for testing population stability

Several amendments to the standard PSI methodology are proposed below. The motivation for each amendment is discussed heuristically before the the proposed methodology is stated formally.

The main costs associated with using an outdated model are divided into two categories. The first is associated with underestimating the number of defaults in a given risk bucket. In this case, the loss incurred by the credit provider is the result of pricing the loan (or other financial instrument) under false, overly optimistic assumptions relating to the number of defaults. The second cost is the opportunity cost associated with clients choosing to take their business elsewhere due to an overpricing of the product that they are interested in. This will be the case when the price of the product is calculated under the assumption that the number of defaults will be unrealistically high.

The first cost mentioned above is explicit and can be quantified with relative ease, while the second cost is implicit. However, both costs are nullified when the number of defaults in each of the classes is modelled accurately by the scorecard used. The expected number of defaults in a given risk bucket is simply the number of clients in that risk bucket taking up the product multiplied by the probability of default associated with this bucket. When testing the hypothesis of population stability, practitioners are typically uninterested in the probability of default for a client falling within a given risk bucket as the probability of default is modelled using separate models. It is the proportion of the population falling into each risk bucket that is of primary concern.

For the reasons explained above, a change in the proportion of the population falling into at least one of the risk categories may result in a loss for the credit provider. As a result, the use of a discrepancy measure based on a maximum discrepancy between the observed and expected differences may be useful. The discrepancy measure employed below is the maximum of the absolute values of the relative differences between the observed and expected proportions in each risk bucket. This type of discrepancy measure has an additional advantage in that it lends itself to a simple, heuristic interpretation in terms of the maximum difference normalised by the size of the risk bucket. Taking this into account, not only is a different test statistic proposed below, but the hypothesis of population stability is reformulated.

The usual formulation of the hypothesis of population stability, given in (1), is quite restrictive. The aim is to test whether or not the population proportion of clients in each risk bucket is exactly as specified under the null hypothesis. In the statistical literature, a hypothesis of this kind is known as a simple hypothesis, meaning that the population is completely specified under the null hypothesis. Below, it is suggested that the hypothesis be restated to enquire whether or not a "material" deviation from the specified model has

been observed. This means that the newly proposed null hypothesis specifies a range of "acceptable" distributions for which the null hypothesis should not be rejected. This kind of hypothesis is known as composite. Below, the critical values of the test are considered and a restatement of the hypothesis to be tested is provided. Thereafter, a more intuitive test statistic is proposed which is directly linked to the (restated) hypothesis of population stability.

The determination as to whether or not a population has remained stable is usually based on a large sample. As a result, if a simple hypothesis (such as the hypothesis in (1)) is tested using a consistent goodness-of-fit test, then very small deviations from the null hypothesis can be detected with a probability close to 100%, as is demonstrated in the examples presented in (4), (5) and (6); see the estimated powers in Table 3. Since the aim of testing this hypothesis is not to detect small changes of this type, the authors advocate for the use of a composite null hypothesis specifying a range of "acceptable" distributions.

In the current context, a "constant of materiality"may be specified; this constant (denoted by δ) determines the range of possible distributions for which the population is deemed to be stable (larger values of δ correspond to a wider range of acceptable distributions). The value of δ is specified according to business considerations associated with the specific scorecard at hand. The value of δ may depend on a number of factors. For instance, in the case where the scorecard is used as a model for a very price sensitive product and the institution in question has a large exposure to the business associated with the product, δ will be chosen small. This is due to the fact that, under these circumstances, a small change in the distribution may have substantial financial implications.

A second proposed change is that not all risk buckets necessarily be included in the calculation. Scorecards are often used in order to determine which clients are eligible for a specific product (such as a loan). As a result, a shift in the proportions associated with the risk buckets to which the product in question is not offered is not of particular interest and should not influence the validity of the scorecard. Consider the case where the product is only offered to clients in the first $k^* \leq k$ risk buckets. In this case, the test should only reject the hypothesis of population stability if a material change is observed in the proportions associated with one or more of the first k^* risk buckets.

Taking the arguments above into account, the hypothesis of population stability can be reformulated as follows. Let k^* denote the index of the risk bucket containing the clients with the lowest credit score that qualify for the product under consideration. The hypotheses to be tested are:

$$H_0 : \max_{j\in\{1,..,k^*\}} |p_j - q_j| \leq \delta q_j, \quad \text{versus} \quad H_A : \max_{j\in\{1,..,k^*\}} |p_j - q_j| > \delta q_j, \tag{7}$$

for some $\delta \geq 0$. Note that, in the case where $\delta = 0$, the hypothesis in (7) reduces to the hypothesis in (1). δ can be interpreted as the maximum allowable shift in the proportions associated with the risk buckets considered, relative to the size of the risk buckets under consideration.

The hypothesis is stated in terms of a maximal discrepancy between the observed sample proportions, $(P_1, \ldots, P_{k^*})$, and the corresponding proportions specified by the null hypoth-

esis, $(q_1, \ldots, q_{k^*})$. A natural test statistic presents itself in order to test the hypothesis;

$$T(\mathbf{q}, \mathbf{P}) = \max_{j \in \{1,..,k^*\}} \frac{|P_j - q_j|}{q_j}. \tag{8}$$

The hypothesis in (7) is rejected for large values of T. Some heuristics as well as the calculation of critical values associated with T are examined below.

Of course, the hypotheses in (7) as well as the test statistic in (8) can be formulated in various ways depending on the needs of the credit provider. For example, the methodology detailed below easily extends to the case where a different constant of materiality is chosen for each risk bucket. If this constant is a decreasing function of the probability of default associated with the risk bucket, then the test will be more tolerant of changes in the proportions of risk buckets with lower probabilities of default. Furthermore, the hypothesis can be changed to define population stability in terms of absolute differences instead of absolute relative differences. The exposition below is not meant to consider every possible situation, it is meant to be a simple, illustrative example which can be altered to suit the needs of the credit provider.

Consider the following example in order to illustrate the behaviour of the newly proposed test statistic under the null hypothesis. Consider the model specified in (3) and fix the sample size at 10^5. Assume that the product in question is offered to the clients obtaining a risk rating in the top 60% of the population (in this case this corresponds to the first six of ten risk buckets). Fix the constant of materiality at $\delta = 20\%$. This means that the maximum tolerated deviation from the model in each risk bucket is 20% of the size of the bucket specified by the scorecard (which corresponds to a maximum absolute difference of 2% in each risk bucket since these are specified to be of equal size). Consider two possibilities for the observed proportions in each risk bucket; let

$$\mathbf{b_6} = \mathbf{q} = (0.1, \ldots, 0.1)$$

and let

$$\mathbf{b_7} = (0.08, 0.12, 0.08, 0.12, 0.08, 0.12, 0.08, 0.12, 0.08, 0.12).$$

Both $\mathbf{b_6}$ and $\mathbf{b_7}$ satisfy the hypothesis of population stability; note that $T(\mathbf{q}, \mathbf{b_6}) = 0$ and $T(\mathbf{q}, \mathbf{b_7}) = 20\%$. However, these two configurations represent two extremes. The first corresponds to an exact match between the specified model and the observed sample, while the latter corresponds intuitively to the configuration satisfying the hypothesis of population stability which is "least like" the specified model. It is possible to sample from the distribution of T using steps 1 to 4 of Algorithm 2 (when one replaces the calculation of $\Psi(\mathbf{q}, \mathbf{b})$ with that of $T(\mathbf{q}, \mathbf{P})$).

Figure 4 shows density estimates of the distribution of T under $\mathbf{b_6}$ and $\mathbf{b_7}$ using solid and dashed lines, respectively. These density estimates are based on 1 million Monte Carlo simulations in each case. The densities shown in Figure 4 shows that the true discrepancy between the current distribution and the model has a substantial effect on the distribution of the values of T.

As was done for the PSI, one can estimate the quantiles of the distribution of T using Monte Carlo simulation. Consider, for example, the case where the sample size is varied;

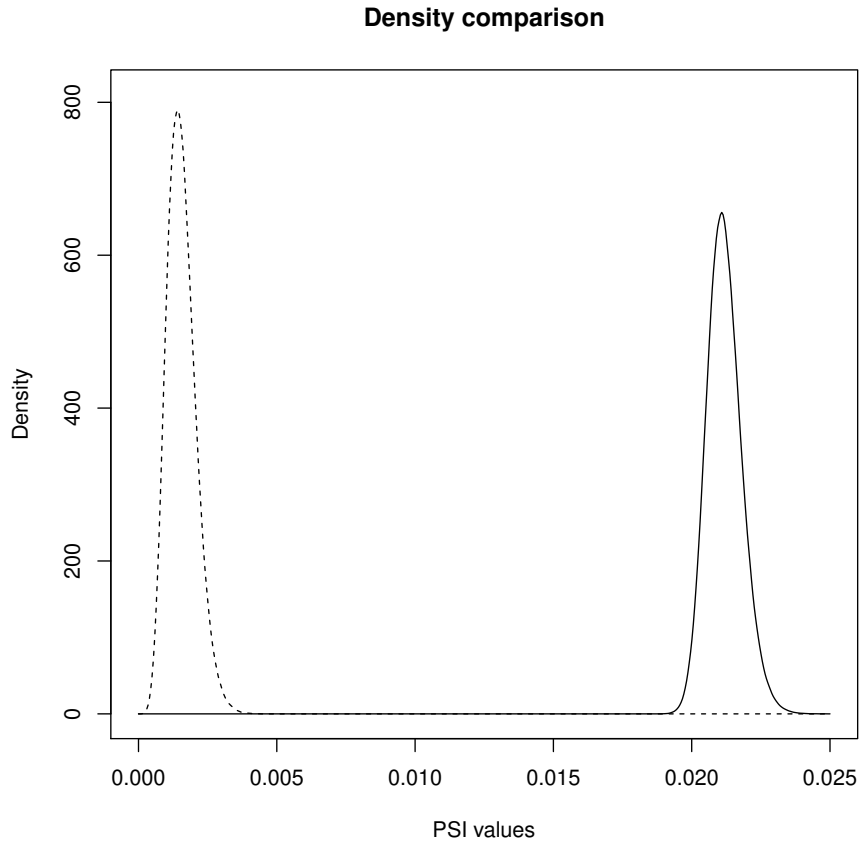

Figure 4: *Kernel density estimates of the distribution of T under* $\mathbf{b_6}$ *(solid line) and* $\mathbf{b_7}$ *(dashed line), respectively.*

$n \in \{10^3, 10^4, 10^5, 10^6\}$, the number of risk buckets is varied; $k \in \{10, 20\}$ and $k^* = 0.6k$. These configurations are realistic values for the South African credit market. Using the Algorithm described above, one may obtain estimated critical values at any given level of significance. Table 3, below, shows the estimated critical values obtained in this way, each table entry is based on 1 million Monte Carlo replications.

Sample size	$\mathbf{10^3}$	$\mathbf{10^4}$	$\mathbf{10^5}$	$\mathbf{10^6}$
$k = 10$	43.0%	27.2%	22.3%	20.7%
$k = 20$	40.0%	26.2%	21.9%	20.6%

Table 4: *95% quantiles of the distribution of T for various sample sizes and numbers of risk buckets.*

Note that the hypothesis of population stability is not rejected for any of the examples $\mathbf{b_3}$, $\mathbf{b_4}$ or $\mathbf{b_5}$ using the newly proposed methodology. The critical value is a decreasing function function of the sample size. Table 3 suggests that, for large samples, one may simply compare the observed value of T to that of δ and conclude that the hypothesis of

population stability is rejected if, and only if, $T > \delta$. However, for smaller sample size, it is recommended that the critical values (as calculated above) be used.

4 Conclusions

This paper is concerned with testing the hypothesis of population stability in credit risk scorecards. The statistical properties of the population stability index are examined numerically for sample sizes typically found in the South African credit market. The properties of the population stability index, together with the lack of simple interpretation available when using this statistic, motivates research into an alternative methodology for testing the hypothesis of population stability. As a result, an alternative specification of this hypothesis is proposed together with an intuitive goodness-of-fit statistic that can be used to test this hypothesis. An algorithm is provided in order to obtain critical values for this test statistic and a table containing calculated critical values is provided for configurations often found in the South African credit market.

5 Acknowledgement

The authors are grateful to Prof Leonard Santana for assistance with the typesetting of the paper.

References

[1] Fan D., 2020, *Creditmodel: Toolkit for Credit Modeling*, R package version 1.1.9, Available from https://CRAN.R-project.org/package=creditmodel.

[2] Hlavac M., 2018, *Stargazer: Well-Formatted Regression and Summary Statistics Tables.*, R package version 5.2.1, Available from https://CRAN.R-project.org/package=stargazer.

[3] Kullback S., 1959, *Information Theory and Statistics*, John Wiley & Sons Inc, London, England.

[4] Kullback, S., and Leibler, R. A., 1951, *On Information and Sufficiency.*, The Annals of Mathematical Statistics, **22(1)**, pp. 79–86.

[5] Meintanis, S. G., 2016, *A review of testing procedures based on the empirical characteristic function.*, The South African Statistical Journal, **50(1)**, pp. 1–14.

[6] Pruitt R., 2010, *The Applied Use of Population Stability Index (PSI) in SAS®Enterprise Miner™Posters.* SAS Global Forum, Available from http://support.sas.com/resources/papers/proceedings10/288-2010.pdf.

[7] R Core Team, 2019, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Available from https://www.R-project.org/.

[8] Siddiqi N., 2006, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons Inc, Hoboken, New Jersey.

[9] Siddiqi N., 2016, *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*, Second Edition, John Wiley & Sons Inc, Hoboken, New Jersey.

[10] Yurdakul, B, 2018, *Statistical Properties of Population Stability Index*, PhD thesis - Western Michigan University, Available from https://scholarworks.wmich.edu/dissertations/3208.