"ALL MODELS ARE WRONG, BUT SOME ARE USEFUL" THE MODEL RISK OF CREDIT SCORING MODELS

Zoltán Pollák – Ádám Kocsis

ABSTRACT

In finance, models are used to support decision-making because observing reality in its totality is impossible. The 2008 crisis brought the flaws of applied models sharply into focus and highlighted the importance of model risk. This paper specifically seeks to quantify the model risk of credit scoring. To this end, the authors begin by presenting the method for determining possible portfolio-level losses caused by model flaws. Using as much information as possible gauged from the tails of loss distribution, the different extents of model risk can be determined by means of the so-called extreme value theory. Following a review of relevant theory, the described procedures are demonstrated on a publicly available database, with the help of R.[1]

JEL codes: C01, C13, C19, C25, C52, C58, G21, G28
Keywords: model risk, credit scoring, reject inference, extreme value theory, value at risk (VaR), expected shortfall

1. MODEL RISK

Processes occurring in business life are impossible to describe in every detail on account of their sheer complexity. Consequently, we create models that enable us to systematise and condense our knowledge. It follows from this simplification that we must always remember our model is merely a scaled-down version of reality, and certainly not equal to it. In addition, our conclusions drawn on the basis of the model are only valid locally, within a given set of circumstances, making it all the more essential that our assumptions should be met.

The crisis of recent years threw the flaws of previously used models sharply into focus, and in particular highlighted the importance of managing model risk. This paper takes its title from a quotation from *George E. P. Box*. Prior to the outbreak

1 This study was created under the research and development project "Research into innovative mathematical models for measuring Basel bank risks and the quantification of capital requirements in the area of market, operational, liquidity and secondary risks; and the behaviour-based prediction of the price trends of financial products" (Project no. PIAC_13-1-2013-0073), funded within the framework of the New Széchenyi Plan, with support from the European Union.

of the crisis, the British statistician, who died in 2013, very aptly pointed out the significance of model risk when he said: "Essentially, all models are wrong, but some are useful" (*Box–Draper*, 2007, p. 414).

We therefore need useful models. However, because models can prove wrong on account of their simplifying nature, particular emphasis needs to be placed on the model risk. Organisations regulating the financial sector have become aware of this factor and expect institutions under their supervision to take it into account when assessing risk.

1.1. A brief summary of regulation

The second pillar of the Basel II Accords requires credit institutions and investment firms to undertake an Internal Capital Adequacy Assessment Process (ICAAP), whereby an institution must assess its own risk profile. Since the circle of risks is broader than under the first pillar, in most cases firms need to meet increased capital requirements compared to the regulatory minimum. However, where the capital requirements calculated by the ICAAP are lower compared to the Pillar 1 requirements – and the National Bank of Hungary (MNB) has approved this by means of a Supervisory Review and Evaluation Process (SREP) – the financial institution is not required to increase its capital.

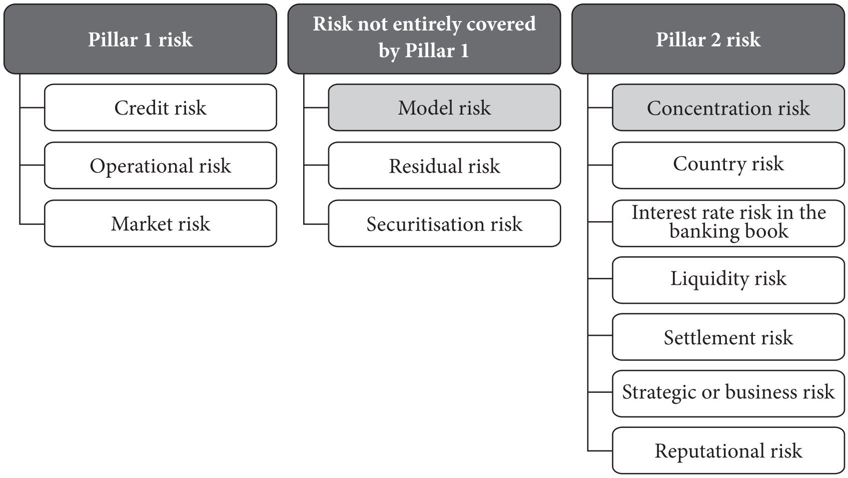
Under Pillar 2, the Basel Committee on Banking Supervision wishes to encourage firms to more consciously measure their own risks and apply more modern and accurate risk management tools. Integrated into various processes, the knowledge gained in this way can contribute to the prudent management of an institution – which is not only a supervisory requirement, but also in the interests of every stakeholder.

The regulator expects credit institutions and investment firms to assess, as a minimum, the following risks:

This paper focuses on model risk, which, as *Chart 1* reveals, was not entirely accounted for under the first pillar. Consequently, institutions must review the correct way of dealing with model risk and perform their internal capital calculations accordingly.

Model risk "is the risk that the institution makes decisions (e.g. in assessment and valuation) that result in financial losses due to model deficiencies" (ICAAP Guidelines, 2012, p. 25). The guidelines point out that the underlying primary cause of model errors is not necessarily negligence, but knowledge limits, insufficient data or changes which cannot be predicted from historic data. This type of risk arises simply from the fact that models are never perfect.

Chart 1
The position of model risk among other bank risks



Source: chart designed by authors, based on the MNB's methodological guidelines for the Supervisory Review Process (SRP)

It is extremely difficult to quantify model risk. Model flaws can be estimated by means of stress and sensitivity tests; however, converting them to losses is possibly an even more difficult matter. Bearing this in mind, it is not so much the case that the regulator expects firms to hold additional capital, but more that it recommends procedural measures to the supervised institutions.

Models are used in many areas in finance. Next, we shall focus on a subsection of model risk: the risks associated with credit scoring models.

1.2. Credit risk and scoring models

Traditionally, credit institutions receive deposits and grant loans. However, while on the one hand a bank will receive deposits almost without restriction, on the other hand it will be very choosy about whom it lends to, and under what conditions.

To offset any possible losses to their credit portfolios, banks are required, under the first pillar of Basel II, to accumulate capital. Credit risk "refers to the risk that a borrower will partially or completely default on a debt by failing to make required payments when they are due" (*Radnai–Vonnák,* 2010, p. 14).

The significance of the problem is highlighted by the fact that banks assign two-thirds (and frequently up to three-quarters) of their capital to cover credit risk, which makes credit risk the most significant bank risk (*Krekó,* 2011). This goes to show that comprehensive management of the problem is a cardinal issue at financial institutions.

There are numerous ways of preventing losses in banking practice. One such method is when a bank applies limits to the amount of credit it will grant to certain institutions or sectors. This helps to avoid credit risk resulting from over-concentrated lending. Further risk-reducing methods include the demanding of guarantees which can be sold when borrowers default on their loans.

The most basic way to manage credit risk, however, is for the bank to make preliminary assessments, as effectively as possible, of whether a borrower will be able to repay the loan plus the interest thereon. Credit scoring models help to distinguish between – or in other words, classify – potentially good and bad clients.

Credit rating is in fact coeval with lending. However, until the first half of the 20th century, loan applications were considered purely on a professional basis, without the use of statistical tools. A major breakthrough occurred in 1941, when *David Durand* first applied a scoring system based on discriminant analysis to private individuals applying for automobile loans (*Kiss*, 2003).

Evidently the most widespread statistical method today is the logistic (or logit) regression model, which was first proposed by *Delton L. Chesser* in 1974 to predict a borrower's default probability.

Since then numerous other models have emerged that are able to effectively deal with the problem of classification without requiring any preliminary assumptions regarding the statistical population. It is the very automated nature of these models that poses the greatest danger since they may have a tendency to function like black boxes, and those applying them often fail to consider the potential dangers.

In the following, we shall attempt to identify the shortcomings of credit scoring models where model risk occurs. The first such cardinal issue is that of the representativity of basic data.

2. MODEL RISK AS A PROBLEM OF MISSING DATA

With respect to model risk, the fundamental problem lies not in the credit scoring models, but in the basic data themselves. No matter how sophisticated the model we use to decide to which client we lend money, it will fail if the data used for modelling are inadequate. The problem with samples corresponds to the phenomenon which the literature describes as selection bias (*Little–Rubin*, 2002).

Selection bias occurs because the sample used for modelling is not, generally speaking, representative. This is because values can only be assigned to every variable in the case of clients who have already undergone a selection process (i.e.

they have already been granted credit). In the case of clients who have not been granted a loan, we have no information regarding whether or not they would have met their payment obligations.

In a fictitious institution where applicants were to be granted a loan by tossing a coin, on a heads-or-tails basis, it would be reasonable to assume that the distribution of variables between clients who are granted a loan is identical with that of rejected clients; that is, the sample is representative of the entire population. In practice, however, banks employ a variety of models to predict if a client will be good or bad (i.e. if they will default on their loan). Consequently, acceptance of an application is not performed randomly; therefore borrowers (whose data we use to construct a model) will not be representative of all applicants.

Also, selection bias is generally a wrong-way risk. To demonstrate this, let us follow the above train of thought. The generally subtle credit scoring models of banks are thought to ensure better selection than credit scoring on a random heads-or-tails basis. In this case, it is reasonable to believe that good clients represent a larger proportion among accepted applications than among rejected applications. Consequently, the model will be constructed from a sample where good clients are over-represented, and the fact that proportionally fewer bad clients make it in will mean that the model will less reliably predict the attributes of bad clients than if the sample was representative of the entire population and included a proportionally larger number of bad clients.

Ignoring the problem will degrade the classification ability of the credit scoring model; and model errors (that is, misclassification) could cause losses for the credit institution.

The literature proposes various techniques to remedy selection bias. Commonly referred to as reject inference[2], these techniques involve the inclusion of rejected clients in the model; for example, in a way that provides a prediction of their behaviour if they had been granted a loan.

2.1. Types of missing data

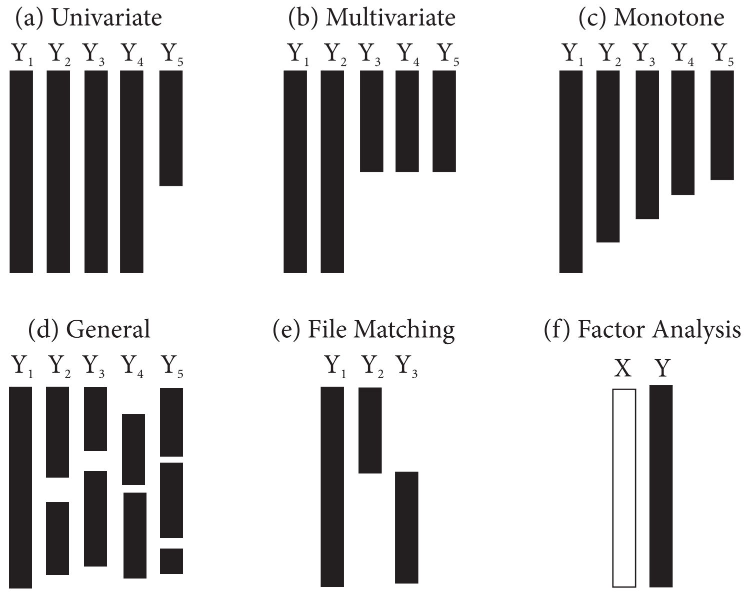
Dealing with missing data is a relatively new area of statistics. The first efforts seeking to comprehensively deal with the problem emerged in the early 1970s, in the early days of computing. In the following, we shall take a brief look at the basic types of missing data so that the terms and concepts to be introduced can be subsequently used in analysing the model risk aspects of the problem.

One approach to describing missing data involves attempting to determine its patterns (*Little–Rubin*, 2002). Traditionally, when building a credit scoring model,

2 The term commonly used to describe methods aimed at reducing selection bias.

the data matrix will have individual observations in the rows and the examined variables in the columns. The variables might be duly completed (which is obviously the ideal scenario), or data relevant to individual units of observation might be missing. Depending on how the missing values are distributed across the basic data matrix, we can distinguish between six different patterns of missing data.

Chart 2
Examples of missing-data patterns*



Note: *The coloured-in part for every variable (column) indicates the observed values.
Source: Little–Rubin (2002)

Case (a) in *Chart 2* is where only one variable has missing data, i.e. where the missing values pose a problem with respect to a single variable, while other variables are complete. The issue explored in this paper can typically be described by this pattern, since in our case a single variable – the explanatory variable that embodies credit risk – is missing values (for rejected clients), and the other variables describing the attributes of potential borrowers assign values to every observation. Naturally, the assumption is that when applying for a loan, all applicants provided all information required by the bank. Let us therefore focus on the case where only one variable has missing data.

In a different approach, missing data can be best measured and dealt with where some knowledge is available of the relationship between the missing elements and the individual variables, i.e. where the process that has caused missing data is known. Cases can be divided into three groups – missing-data mechanisms – depending on the degree of randomness of the missing data (*Oravecz*, 2008):

– Missing Completely at Random (MCAR)
– Missing at Random (MAR)
– Not Missing at Random (NMAR)

The data matrix can be written as $Y = (y_{ij})$ containing n number of observations according to K number of variables. Let us introduce an $M = (m_{ij})$ indicator matrix the value of whose m_{ij} elements equals 1 where data is missing and 0 where the data is observed. Formally, the nature of the missing data can be described by the conditional distribution of M given Y ($f(M|Y, \theta)$, where θ refers to unknown parameters (*Little–Rubin*, 2002).

We say that the data is Missing Completely at Random (MCAR) when the distribution of fully and partially observed individuals is identical, that is, in the above-defined conditional distribution matrix M does not depend on Y:

$$f(M|Y,\theta) = f(M|\theta) \text{ is ful filled in the case of } \forall Y, \theta . \tag{1}$$

An example of such a missing-data mechanism would be where a bank was to decide on a heads-or-tails basis whether or not to grant a loan to an applicant.

We say that the data is Missing at Random (MAR) when the missing data cannot be inferred from the missing variable, but can be predicted by means of the other (complete) variables.

$$f(M|Y,\theta) = f\left(M \middle| Y_{observed}, \theta\right) \text{ is ful filled in the case of } \forall Y_{missing}, \theta, \tag{2}$$

where $Y_{observed}$ is the component of matrix Y containing complete observations, while $Y_{missing}$ is the part where the missing data occur. The data missing at random that corresponds to equation (2) is illustrated by the following case:

Let us assume we possess a sample of credit applications fully completed by the clients, with no missing data. Next, we build a credit scoring model, on the basis of which we decide which clients we will offer a loan. After lending the money we observe which clients have met their obligations and which clients have defaulted. In the case of the latter parameter, which represents credit risk, we shall naturally discover missing values among the rejected clients; however, since we decided who should be granted a loan on the basis of a straightforward, well-documented method, the missing data can be inferred from the other (complete) variables on account of the fact that our credit scoring model was built with the same complete variables.

It has to be stressed that in the above example the credit scoring procedure was performed according to clearly determined rules. If we were to enter ad hoc elements into the selection algorithm (by allowing exceptions/overrides), missing data could not be inferred from the other variables and consequently we would slip into the next, considerably less favourable category.

The next category is that of data Not Missing at Random (NMAR), meaning that the missing data of the incomplete variable cannot be inferred from the other variables. This scenario is the most difficult one to deal with of all types of missing data.

Identifying missing-data mechanisms is crucial to adequately dealing with the problem and quantifying the resulting risk (or uncertainty).

2.2. A logistic regression imputation model based on a Y_i dichotomous variable

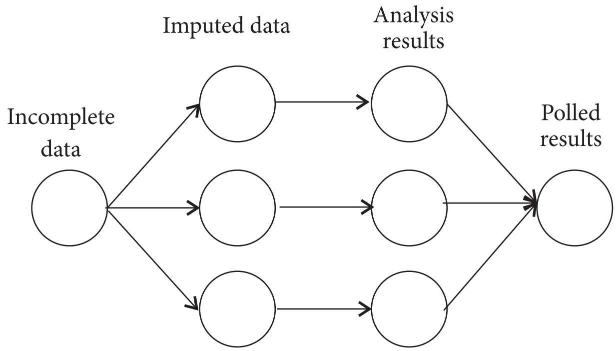
In the above we have sought to point out that the phenomenon of selection bias is an existing problem in credit scoring models, and a significant one at that. There is no universal method for reducing the distortion; we need to consider various aspects before choosing any particular method.

By attempting to estimate non-existing data in an incomplete database, we unwittingly enter uncertainty into our estimation. We have elected to use a logistic regression-based multiple imputation model because it enables an approximation of the variance of estimators; also, the uncertainty caused by missing data can be incorporated into the system (*Oravecz*, 2008).

In multiple imputation models, “multiple” refers to the fact that for every missing value an m number of estimations are made, and at the end we pool the results of the analysis performed on the complete m database with the help of the estimated parameters and standard errors (*Little–Rubin*, 2002). The uncertainty of substitutions is incorporated into the model, so that the imputed database is able to approximate the variability of the complete database.

The logic of multiple imputation and the above-described main steps are illustrated in the following figure ($m = 3$):

Chart 3
The main steps of multiple imputation



Source: Buuren–Groothuis-Oudshoorn (2011)

The three iterations in *Chart 3* would be too few in practice, but 10–20 would suffice in most cases (*Buuren–Groothuis-Oudshoorn*, 2011).

Our model involves a multiple imputation procedure that is applicable in the case of single-variable missing data where the incomplete variable is dichotomous (its values being exclusively 0 or 1).[3]

Let θ denote a vector containing unknown parameters, and X_i the set of completely observed variables (these will be the explanatory variables). Let Y_i be the incomplete dichotomous variable whose missing values we endeavour to estimate. The conditional distribution of dummy variable Y_i will be as follows:

$$f(Y_i|X_i,\theta) = logit^{-1}(X_i\theta)^{Y_i}[1 - logit^{-1}(X_i\theta)]^{1-Y_i}\ , \qquad (3)$$

where the inverse of the logit function means the following:

$$logit^{-1}(a) = \frac{e^a}{1+e^a}\ . \qquad (4)$$

The estimation of the θ parameters is performed by means of the maximum likelihood method on the basis of the data of the fully observed (that is, the accepted applications). The procedure accomplishes the imputation of the missing values in three steps.

(1) First, we draw from the normal distribution, whose expected value and variance is calculated from the data of the observed variables, as many random numbers as we have X_i explanatory variables (let θ denote the resulting vector).

3 For a detailed description of the procedure, see *Rubin* (1987), p. 169.

(2) Second, with the help of this estimated parameter vector, we calculate the value of $logit^{-1}$ (X_i θ_*) for every missing observation, which will yield a real number between 0 and 1.

(3) Third, we generate for every missing value a uniformly distributed random number on the interval 0, 1 (denoted as v_i). Next, all that needs to be done is replace the missing data with $Y_i = 0$ provided $v_i > logit^{-1}(X_i\ \theta_*)$ and $Y_i = 1$. Given that we are dealing with multiple imputation, we keep repeating these three steps to create an *m* number of independent, complete databases (generating new random numbers each time) whose results are summarised as illustrated in *Chart 3*.

The above-described procedure is a relatively simple multiple imputation method that can be used to deal with single-variable MAR-type missing data, where the incompletely observed variable is dichotomous. *Chapter 4* will present the procedure on a real database, with a view to reducing selection bias.

3. MEASURING THE MODEL RISK OF CREDIT SCORING MODELS

The previous chapter discussed the phenomenon of selection bias, as well as a specific technique aimed at reducing it. This was necessary firstly because the method can help improve the classification ability of credit scoring models, and secondly because in measuring model risk it will be assumed that the results of the logistic regression scoring model can be interpreted as a client's probability of default. The latter assumption will only be met where the model-building sample is representative of the entire population of credit applicants, which could not be accomplished without incorporating the data of rejected clients.

3.1. Classification and the possible losses of scoring models

This section will explore the estimation of model risk in credit scoring models. In order to determine the extent of risk, we firstly need to know the possible losses resulting from the mistakes in credit scoring models.

Two mistakes can be made in the course of classification, each of which entails different costs (*Thomas* et al., 2002). Firstly, we might reject a potentially good applicant (error of the second kind), causing the bank to lose the profit that might have been made on that client. Secondly, the system might classify a potentially bad client as good, resulting in the bank granting the loan (error of the first kind). If the client defaults and fails to fulfil their payment obligations, the bank will suffer actual losses.

A 2 × 2 matrix, the so-called confusion matrix, is drawn up for the comparison of the state predicted by the model and the actual outcome (the applicant defaulted or paid off the debt).

Table 1
Confusion matrix with type I and type II errors

Confusion matrix		Actual Good (G)	Actual Bad (B)	Σ
Predicted	**Good (G)**	True positive (g_G)	False positive (g_B) **type I error**	***g***
Predicted	**Bad (B)**	False negative (b_G) **type II error**	True negative (b_B)	***b***
Σ		$\boldsymbol{n_G}$	$\boldsymbol{n_B}$	***n***

Source: Thomas et al., (2002)

The confusion matrix is built in a way that its rows contain, for a given *C* cutoff (the threshold value for accepting/rejecting clients), the number of applicants considered by the model to be good (*g*) or bad (*b*). With the help of the variable indicating credit risk (i.e. the true nature of the client), we can see how many individuals were classified correctly and how often the classification was mistaken. For a sample of *n* elements, the values listed in the last, summary row of the confusion matrix (n_G, n_B and *n*) are fixed, while the number of accepted (*g*) and rejected (*b*) applicants would depend on the choice of cutoff value.

The structure of the confusion matrix and the number of errors with a given probability of default depends on the choice of cutoff value. The following two tables demonstrate some extreme cases:

Tables 2 and 3
Confusion matrices with cutoff values of 0 and 1

Cutoff = 0		Actual Good	Actual Bad	Σ
Predicted	**Good**	0	**0**	**0**
Predicted	**Bad**	$\mathbf{n_G}$	n_B	**n**
Σ		$\mathbf{n_G}$	$\mathbf{n_B}$	**n**

Cutoff = 1		Actual Good	Actual Bad	Σ
Predicted	**Good**	n_G	$\mathbf{n_B}$	**n**
Predicted	**Bad**	**0**	0	**0**
Σ		$\mathbf{n_G}$	$\mathbf{n_B}$	**n**

Source: original tables

In the first case (left table) the chosen cutoff value was the minimal 0. In this case, every applicant is rejected, resulting only in a type II error, its extent corresponding to the number of good clients in the database (due to the fact that their applications, too, were rejected).

The table on the right illustrates the situation where the rejection threshold is set at the maximum value of $C = 1$. In this case, the bank will offer a loan to every applicant. This case is the oft-quoted open-doors method. It is the best way to avoid the above-discussed selection bias in that it affords a model-building sample that truly represents the entire population. If it is so effective, why then do banks not use it in building their credit scoring models?

The answer is simple: because "experience is an expensive school."[4] Due to defaulting clients the bank would suffer immeasurable losses on such a credit portfolio, so instead it goes along with a larger number of model errors caused by selection bias, or seeks a less reliable solution regarding the representativity of basic data.

A cutoff is, then, a variable whose adjustment causes the structure of the confusion matrix to change, with a given probability of default. This paper seeks to quantify model risk as the bank's risk of loss due to model errors. Loss due to model error is caused by type I and type II errors; consequently, in the following we need to focus on the top right and bottom left quarters of the 2 × 2 confusion matrix (*Table 1*).

First, by means of one process or another (for example, logistic regression), we estimate $p(x)$ as the conditional probability of default for every client, whose interpretation as a probability is based on whether the model-building sample is representative of the population of clients "walking in from the street." Selection bias can be reduced by means of various reject inference techniques, and by incorporating the data of rejected clients, it is not unreasonable to assume that the sample is representative. The probabilities of default are therefore given, and like cutoff values, they have a value on the interval (0, 1).

If the individuals are arranged in a row according to their estimated probability of default, provided no two observations are identical, the cutoff value between any two adjacent values of $p(x)$ probabilities of default will lead to different confusion matrices. This enables the generation of $(n+1)$ different structures of confusion matrix, which, with given probabilities of default, depend on the observed values of the default variable. In turn, these differently composed confusion matrices allow for the modelling of model risk in credit scoring systems.

Next, let us divide the interval (0, 1) into several partial intervals as discussed above, and take a look at the loss caused by complete model errors, along the individual division points as cutoff values.

Let it be assumed that the cost of a type I error is D (*debt*), which is identical for every loan applicant. For example, $D = 0.45$ would mean that where an accepted

4 *Benjamin Franklin* (quoted by Jorion, 1999, p. 40).

bad applicant defaulted, the bank would fail to recover 45% of its exposure to the given client. Let L (*lost profit*) denote the opportunity cost arising in the event of a type II error, which the bank would suffer on losing interest income. In the course of estimating model risk, we always take the ratio of losses resulting from the model errors D and L to be fixed.

With a C rejection threshold value, the loss caused by model errors can be calculated for the entire credit portfolio as follows:

$$Total\ loss_C = D \sum_{i=1}^{g_B} E_i \cdot \widehat{p(x)}_i + L \sum_{i=1}^{b_G} E_i \cdot \widehat{p(x)}_i \quad , \tag{5}$$

where E_i is the volume of exposure of the i^{th} individual (that is, the sum of credit granted in the absence of a guarantee); and $\widehat{p(x)}i$ is the conditional probability of default of the given client.

In equation (5) the first summation applies to clients in respect of whom the model made a type I error; the second summation combines losses caused by individuals in respect of whom a type II error occurred.

In this section we have demonstrated how to obtain possible portfolio-level loss values caused by the classification model. The next subchapter will show how, using the distribution of losses, different extents of model risk can be determined.

3.2. Measuring model risk with the help of the extreme value theory

Values better than empirical quantiles can be obtained to express the extent of model risk in credit scoring systems, where, by applying the extreme value theory to determine the value at risk (VaR), an adequate distribution is fitted to the tails of loss. This is necessary because of the typical paucity and scarcity of observations in the tails of loss distribution, causing our point estimation to be misleading.

The extreme value theory (EVT) deals with the statistical analysis of extreme events. With respect to the financial applications of the theory, the most widespread model is that of threshold exceedances (or peaks over threshold) which, in estimating the tails of loss, takes into account every loss that exceeds a certain u loss threshold (*Tulassay*, 2013). This enables a better estimation of the VaR by fully taking into account the tails of loss.

Let X be a probability variable that represents the losses to be modelled, and $F(x)=P(X \leq x)$ the loss distribution function. Let extreme loss be regarded as a value that exceeds the u threshold. The distribution of exceedances (assuming the u threshold has been crossed) is then:

$$F_u(y) = P(X - u \leq y | X > u) \quad (6)$$

where $(X - u)$ is none other than the extent of exceedance.

Bayes' theorem can help connect the loss distribution function $F(x)$ and conditional excess distribution function $F_u(y)$.

$$F_u(y) = P(X - u \leq y | X > u) = \frac{P(u < X \leq y + u)}{P(X > u)} = \frac{F(y + u) - F(u)}{1 - F(u)} . \quad (7)$$

The *Pickands–Balkema–de Haan* theorem states that for a large class of distribution functions there exists a ξ and a β(u) where, if the u threshold approaches the upper endpoint of the distribution, the following will be true for the conditional excess distribution function:

$$F_u(y) = P(X - u \leq y | X > u) \approx G_{\xi,\beta(u)}(y), \quad (8)$$

where $G_{\xi,\beta}(y)$ is the generalized Pareto distribution (GPD). The *Pickands–Balkema–de Haan* theorem then states that for an adequately high threshold, exceedances approximate a GPD distribution, making the generalized Pareto distribution a natural model of threshold exceedances (*McNeil* et al., 2005).

The standard cumulative distribution function of the GPD is defined by:

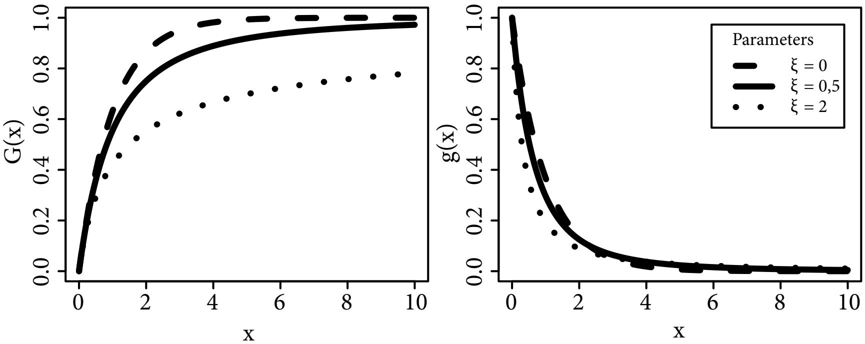
$$G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - e^{-\frac{y}{\beta}}, & \xi = 0 \end{cases} \quad (9)$$

where ξ is the so-called shape parameter, and $\beta > 0$ the scale parameter. The expected value of the GPD distribution is

$$E(X) = \frac{\beta}{1 - \xi} . \quad (10)$$

Definition (9) clearly shows that where $\xi = 0$, the GPD follows an exponential distribution with parameter $\lambda = \frac{1}{\beta}$, that is, exponential distribution can be considered a special case of generalized Pareto distribution. *Chart 4* shows the GPD distribution function (left) and the GPD density function (right) for three different ξ shape parameters (where $\beta = 1$ in every case):

Chart 4
GPD distribution and density functions for different ξ values



Source: original chart, created in R

The dashed line indicates exponential distribution (ξ = 0), the uninterrupted line the GPD distribution for shape parameter ξ = 0.5, and the dotted line the GPD distribution for shape parameter ξ = 2.

As the density functions (g*(x)*) in *Chart 4* reveal, by choosing the right shape parameter the GPD can be flexibly placed on the edge of the distribution, and with the help of the scale parameter it can be applied to absolute losses expressed in monetary units or even returns expressed in percentage.

Above a certain u threshold, using equations (7) and (8) and taking advantage of the fact that $x = y + u$, the model of the tails of loss distribution is as follows ($x > u$):

$$F(x) = [1 - F(u)]G_{\xi,\beta}(x - u) + F(u)\,, \tag{11}$$

where *F(u)* is usually estimated from historical data: $\hat{F}(u) = \frac{n-N_u}{n}$, a method that the literature refers to as historical simulation (N_u in the equation refers to the number of losses exceeding the u threshold, and n the total number of observed losses) (*McNeil* 1999).

In this case $x > u$ losses can be modelled as follows:

$$\hat{F}(x) = \left[1 - \hat{F}(u)\right]G_{\hat{\xi},\hat{\beta}}(x - u) + \hat{F}(u) = 1 - \frac{N_u}{n}\left[1 + \hat{\xi}\left(\frac{x-u}{\hat{\beta}}\right)\right]^{-\frac{1}{\hat{\xi}}}\,, \tag{12}$$

which is a more specific model than if we were using empirical distribution only.

In describing the model of threshold exceedances, that well-chosen u threshold comes up often, and the theory models excesses above the threshold. In practice, determining the value of u is not an easy matter. One possible way of choosing the threshold is to examine the mean excess function. The mean excess function of probability variable X (where the expected value of X is finite) would be as follows:

$$e(u) = E(X - u | X > u) \ . \tag{13}$$

As we have mentioned, $F_u(y)$ in equation (7) is the distribution of exceedances of the u threshold, assuming that the loss exceeded the given threshold. The mean excess function in equation (13) provides the expected value of $F_u(y)$ as a function of u. The fact that the conditional excess distribution follows the GPD distribution, that is $F_u(x) = G_{\xi,\beta(u)}(x)$, is conditional on $\beta(u) = \beta + \xi u$. Using the expected value of the GPD distribution in equation (10), the mean excess function can be converted as follows:

$$e(u) = \frac{\beta(u)}{1-\xi} = \frac{\beta + \xi u}{1-\xi} = \frac{\beta}{1-\xi} + \frac{\xi}{1-\xi} u \ . \tag{14}$$

Equation (14) reveals how in the case of the GPD distribution the mean excess function is linear in u, meaning that when choosing the adequate threshold we need to plot $e(u)$ as a function of u and find a threshold value above which the function is approximately linear; since in this way the GPD is expected to fit well.

The above-described model of threshold exceedances uses the information in the tails of distribution better than the simple quantile of loss distribution, making it suitable for estimating the VaR more accurately than before, as follows:

$$\hat{F}(x) = 1 - \frac{N_u}{n}\left[1 + \hat{\xi}\left(\frac{x-u}{\hat{\beta}}\right)\right]^{-\frac{1}{\xi}} = q \, . \tag{15}$$

The VaR can be obtained by inversion of the function ($q>F(u)$):

$$\widehat{VaR}_q = F^{-1}(q) = u + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{n}{N_u}(1-q)\right)^{-\hat{\xi}} - 1\right]. \tag{16}$$

The obvious advantage of VaR lies in its simple interpretability; though it does have a number of disadvantages, including the fact that it cannot be regarded as a coherent measure. However, the greatest problem with the VaR is not this, but the fact that it says nothing about losses exceeding it; that is, the tail of the distribution containing the greatest losses.

The model enables the calculation of coherent risk measures, such as expected shortfall (ES), for example. ES uses information in the tails of distribution, revealing the (conditional) expected value of losses exceeding the VaR.

$$\widehat{ES}_q = \widehat{VaR}_q + E\left(X - \widehat{VaR}_q | X > \widehat{VaR}_q\right) = \frac{\widehat{VaR}_q}{1-\hat{\xi}} + \frac{\hat{\beta} + \hat{\xi} u}{1-\hat{\xi}} \ . \tag{17}$$

The model of threshold exceedances affords better use of the information in the tails of distribution, as well as more accurate and stable results than the empirical quantile. Beautiful as the theory is, it has the shortcoming that it allows for estimation of very high quantiles with a high rate of error. The estimation of the parameters of GPD distribution, as well as other issues crucial to the topic of this paper, will be demonstrated via a practical example.

4. DEMONSTRATING MODEL RISK ON A SPECIFIC REAL DATABASE

The purpose of this section is to demonstrate the methods reviewed above on a real, publicly available database. Following a brief introduction of the data set, we will first apply a reject inference technique – specifically, a multiple imputation-based process – with a view to reducing selection bias. Based on logistic regression, we will then build a scoring model on the imputed database, whose model risk we will quantify with the help of a coherent (ES) and an incoherent (VaR) risk measure.

4.1. Introduction of the *German Credit Data* data set[5]

Credit scoring data sets and the credit scoring models based on them are the most closely guarded information of any bank. This is why it is extremely difficult to obtain access to a complete data set that can be used to demonstrate the procedures described above. A few smaller data sets are available for educational purposes and one of these will be used to perform the necessary analyses, with the help of the statistical software R.

The *German Credit Data* data set was published by the Institute of Statistics and Econometrics at the Department of Economics of the University of Hamburg. The data set contains the attributes of 1,000 private credit applicants. The rows contain the observed units (clients), the columns the individual variables used to assess credit applications. We have a total of 20 explanatory variables and one default variable indicating credit risk to decide if an applicant will be a good or bad client.

After providing a broad outline of the main characteristics of the data set, as the next crucial step we will seek to reduce selection bias in the basic data by means of the above-described method.

5 The *German Credit Data* is a public data set accessible at: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data), downloaded on 09.10.2014.

4.2. Reducing selection bias

As we pointed out in the theoretical introduction, the literature on selection bias offers numerous methods to remedy the issue. No universal procedure exists that can afford the best solution for all types of missing data. Individual authors have often come to contradictory conclusions because the success of one method or another depends on the characteristics of the specific database.

In reality, banks possess all the required data of every rejected applicant (naturally, with the exception of the variable indicating default), meaning that they possess a sample that truly contains the data of applicants representative of the population who "walk in from the street." No such complete database is available publicly, so we can only assume the *German Credit Data* to be such a data set; that is, one that contains all types of applicants consistent with the population proportion.

In this case, we do not know the value of the default variable of rejected applicants, so they need to be deleted from the data set. To decide which clients were rejected, we will need to run a logistic regression for all observed units. Of the available 20 explanatory variables, the variables significant at the 5% level were singled out by means of a backward-type[6] model selection procedure and only those were included in the model. Eventually, we determined a score for each individual on the basis of a narrow model containing 11 explanatory variables.

Let us assume that on the basis of these estimated values the bank rejected the worst 50 credit applications, and the remaining 950 private individuals were granted the loans they applied for. The reason we have to assume such a high (95%) rate of acceptance is because a higher degree of rejection would leave us with too few bad clients in the sample, which could not accommodate the imputation procedure to be introduced later on. A real data set used for building a model would contain considerably more observed units than this data set of 1,000 units; consequently, a higher rate of rejections would not constitute a problem there.

Next, we deleted[7] the *default* variable of the rejected 50 applicants, because we can only know this information for the accepted clients. If we were to operate exclusively with the data of the accepted 950 applicants, we could expect distorted results due to the presence of the selection bias described above. Consequently, we need to include the data of rejected clients in the analysis. This can be achieved by estimating the missing data of rejected applicants on the basis of the fully observed data of the accepted clients.

6 The backward-type model selection procedure narrows down the model step by step until all included variables become significant (Kovács, 2011).

7 Prior to deletion, we saved the values to another object, with a view to using this later on in examining the efficiency of the applied imputation procedure.

Estimation of the *default* variable of rejected applicants is accomplished by means of a logistic regression-based multiple imputation procedure, with the help of the R package "mice." Multiple imputation is a technique often used for dealing with incomplete data sets with data missing at random (MAR). The procedure described in this paper was thus used to estimate unknown values of the *default* variable. Comparison of the imputed values with the previously deleted real values revealed that the procedure erred in 13 out of the 50 cases. Consequently, it correctly estimated missing observations in 74% of cases.

Next, we performed the imputation with the version of the procedure using a bootstrap approach. Some authors believe that the above-mentioned assumption of normality generally becomes impaired under multiple imputations, leading to distorted estimations of the θ parameters. The research of *White* et al. (2010) has revealed, however, that this can be avoided by means of procedures using a bootstrap approach. To this end, samples of the observed data are taken, on which we repeatedly perform the imputation procedure, saving the θ^* parameters generated in each case. Eventually, drawing from the distribution of θ^* parameters we substitute the missing data.

Multiple imputation using the bootstrap was only incorrect in six cases compared with the original values, making for an efficiency rate of 88%. In all six misclassified cases the method made type II errors; that is, it classified truly good clients as bad. Practical experience shows that losses caused by type I errors are considerably larger than those caused by type II errors. Consequently, not only did the procedure we employed make fewer mistakes, but the mistakes it did make were on the lower-cost side. Bearing this in mind, we substituted the missing data matrix with the estimated results obtained in the second procedure.

4.3. Determining possible model risk losses

Next, we ran a logistic regression (as described in the previous section) on the now complete data set; first with the inclusion of all explanatory variables, and then with a narrower model using variables significant at the 5% level.

Using the logit's estimated parameters, we determined the probability of default (PD) for every client. Interpreting the results as probability might sound like a rather steep assumption, but if the 1,000 applicants in the data set truly represent the entire population and our efforts to deal with selection bias have been successful, it is perhaps not so far-fetched after all.

Quantifying model risk requires an estimated *Loss Given Default* (LGD), which is the share of an asset that is lost when a client defaults. The Basel framework and EU legislation serve as points of reference in estimating this, where, under the *Internal Ratings-Based* (IRB) Approach, a 45% LGD value is assigned to senior exposures without eligible collateral (Article 161(1) of Regulation (EU) No. 575/2013). However, this value does not apply to exposures from loans granted to private individuals; credit institutions and investment firms have to calculate their own LGD for private equity exposures. Our analysis will take this regulatory value as a starting point, but any institution may substitute this input parameter with its own estimated LGD.

Following the train of thought to determine the losses resulting from model risk, let the costs of type I errors be 45% for every individual ($D = 0.45$). Caused by rejecting good clients, type II errors represent alternative costs; that is, lost interest revenues. In the description of the *German Credit Data* data set, Professor *Hans Hofmann* of the University of Hamburg estimates the costs of type I errors to be five times higher than those of type II errors (*Hofmann*, 1994). Accordingly, let type II error losses account for 9% of the exposure ($L = 0.45/5 = 0.09$). At the end of the analysis we will test the estimated losses for sensitivity; i.e. examine their affect on the ultimate extent of risk.

The possible losses caused by model risk are determined as described in section 3.1. First, we created the vector of cutoff values along which the losses will be examined, occurring for the different confusion matrices. Next, at different rejection thresholds, we examined which individuals the model classified as false positives or false negatives.

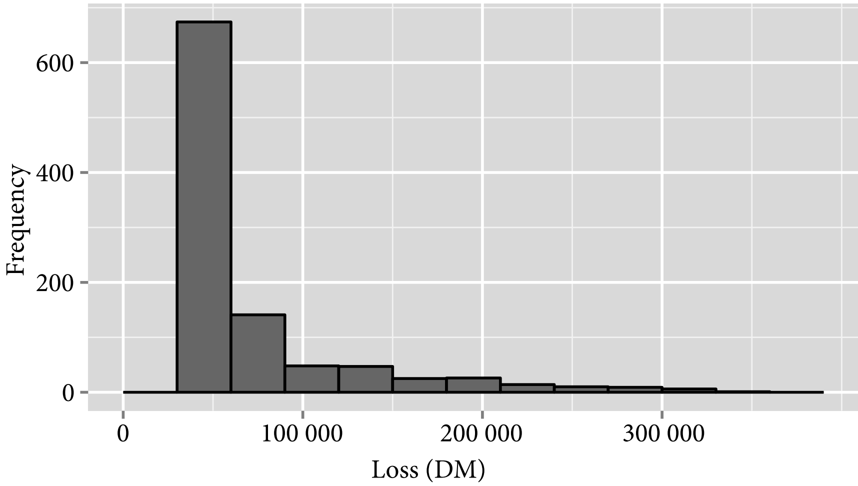
For every applicant that the model misclassified, we calculated the expected loss. This was performed by taking a respective 45% or 9% of the value of the exposure (that is, the amount of the loan), depending on whether the model had made a type I or type II error, which was subsequently multiplied by the probability of default estimated for the given individual. It was assumed here that these risk factors (PD, LGD, ES) are independent of one another, and that we were dealing with exposures without eligible collateral; that is, the extent of the exposure was identical with the amount of the loan.

For a given cutoff, the portfolio-level model risk loss will be the total of these expected individual costs. Calculating this for the acceptance thresholds determined in the manner described above, we obtain the losses caused by model flaws for various confusion matrices.

4.4. Quantifying model risk

The following histogram shows portfolio-level model risk losses caused by misclassification, as calculated above:

Chart 5
Histogram of portfolio-level losses



Source: original chart

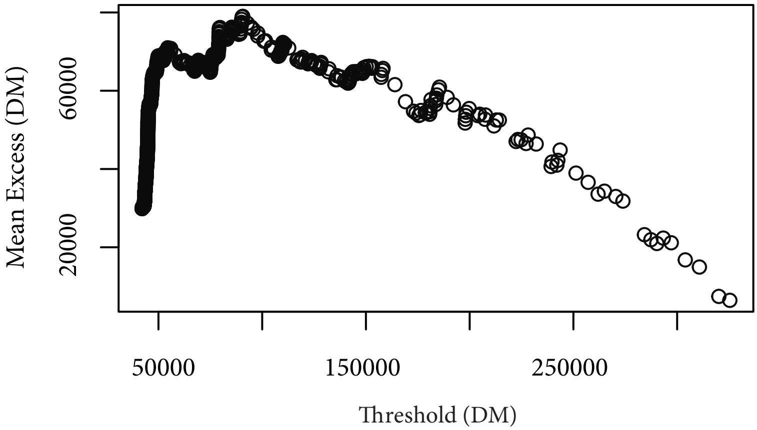
The loss distribution shown in *Chart 5* is highly asymmetrical, with the right tail drawn out. This attests to the fact that low losses occur frequently, while severe model errors are rare.

In the following, we shall apply the model of threshold exceedances to the tails of loss distribution. The calculations and figures were created with the help of the "*evir*" package in R. As we mentioned above, the extreme value theory has numerous areas of application, and the software package R – specifically designed for finance – calculates different extents of risk for loss distribution. The software estimates the parameters of the applied GPD distribution by means of the maximum likelihood method (*Gilleland* et al., 2013).

First, we need an adequately chosen *u* threshold where the conditional excess distribution approximately follows the GPD distribution. The threshold is most often established by means of the mean excess function, which shows the average of losses in excess of various *u* values (horizontal axis).

In GPD distribution this expected value is the linear function of the *u* threshold; in other words, the task is to determine a threshold above which the mean excess function is approximately linear.

Chart 6
Mean excess function



Source: original chart

Chart 6 shows that the problem of choosing a threshold is not necessarily a straightforward one, and that several possibilities exist. Approximately below the 60,000 DM limit the function has a positive slope, while the slope is negative above that value. On this basis, we applied a GPD distribution to the losses for the $u = 60{,}000$ threshold. Because there are several possible thresholds, at the end of the analysis the sensitivity of the extent of risk will be shown for this choice of threshold. For the estimated shape and scale parameters ξ and β, the applied GPD distribution and standard errors were as follows:

Table 4
Estimated parameters and standard errors of the applied GPD distribution

Name	ξ	β
Estimated parameter	–0.0822	72 744
Standard error	0.0532	3 964

Source: original table

Applied to the observations, the GPD model affords the opportunity to estimate the high quantiles of loss distribution, and in turn, to calculate various VaRs.

Table 5
VaR values, historical percentiles and expected shortfall

Significance level	VaR	Historical percentile	ES
95%	5.70%	6.05%	7.46%
99%	8.57%	8.87%	10.12%

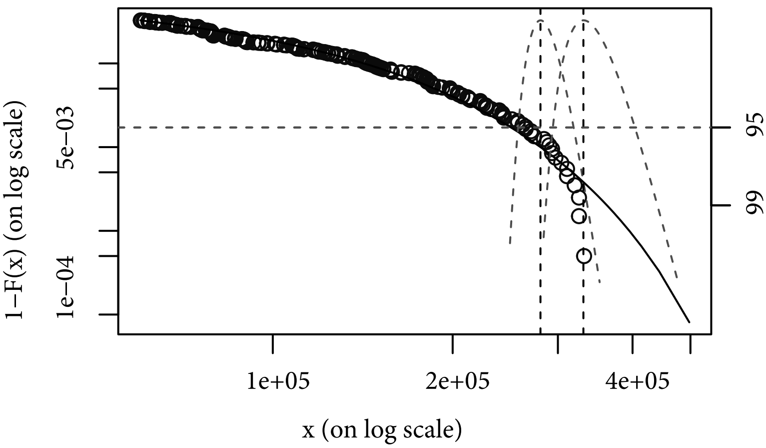
Source: original table

The extent of risk in *Table 5* is expressed in terms of the percentage of portfolio value. It can be seen that the historical percentile of loss distribution is in every case relatively close to the VaR estimate based on the applied GPD distribution. Consequently, the latter can be regarded as the better estimation because, as described above, the model of threshold exceedances better exploits the information in the tails of loss distribution.

The VaR value in the bottom row of *Table 5* can, for example, be interpreted in a way that, at a 99% level of reliability, maximum losses resulting from credit scoring model errors in the next period will be 8.57% of the value of the portfolio. The expected shortfall is the (conditional) expected loss exceeding the VaR, which means it will always be higher than the VaR. Accordingly, the 10.12% value in *Table 5* means that, at a 99% level of reliability, the credit institution or investment firm could suffer a loss of 10.12% in the worst 1% range of possible model risk losses (i.e. in the case of losses exceeding 280,506 DM).

The extent of risks interpreted above, their confidence intervals and the applied GPD distribution are illustrated in the following figure:

Chart 7
VaR and ES values and their confidence intervals at a 99% level of reliability

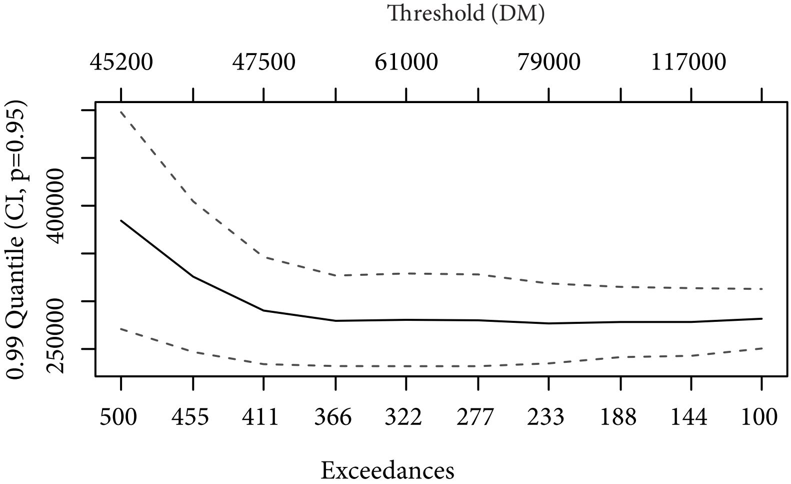


Source: original chart created with the *"evir"* package in R

The circles in *Chart 7* show the individual observed losses (*x*) and the continuous line the applied GPD distribution. It can be seen that apart from a few outliers, the data points more or less follow the applied GPD. The vertical dashed line on the left indicates the 99% VaR value and the vertical straight line to its right the expected shortfall. The two concave curves show the confidence intervals of the extents of risk. The points where the dashed horizontal line and the two curves intersect give the extreme points of the 95% (right axis) confidence interval of the estimated risk values. Moving the line parallel downward towards the 99% value will yield intervals with increasingly greater levels of reliability. Naturally, as can be seen in *Chart 7*, an estimation with a higher level of reliability yields an increasingly broader confidence interval.

As we mentioned earlier, the choice of *u* is not always a straightforward matter on the basis of the mean excess function (*Chart 6*).

Chart 8
The value of VaR at a 99% level of reliability as a function of the u threshold



Source: original chart

Chart 8 shows that above our chosen 60,000 DM limit (top axis) *u* could have been practically any value, as it would not significantly affect the estimation of VaR at a 99% level of reliability. This means that the model of threshold exceedances affords a reliable estimation of the extents of risk (*McNeil et al.* 2005).

Demonstrated on a variety of assumptions5, the described procedure depends on the value of certain input parameters, such as the costs of type I and II errors.

Previously, type I error was determined as *D* = 45% based on the LGD set out under the Basel framework and in EU legislation, while the costs of type II errors were calculated from the error rates with the help of expert estimation (*L* = 9%). The next table shows the sensitivity of VaR (at a 99% level of reliability) to these parameters:

Table 6
The cost effects of type I (D) and type II (L) errors on VaR at a 99% level of reliability

Sensitivity test		D		
		–1%	**0%**	**+1%**
	–1%	–0.91%	**–0.64%**	0.38%
L	**0%**	**–0.73%**	0%	**0.93%**
	+1%	–0.70%	**0.03%**	1.02%

Source: original table

The columns in *Table* 6 show the shift in the value of VaR at a 99% level of reliability when the costs of type I errors are increased or reduced from *D* = 45% in increments of 1%. The rows show the change in percentage of the same extent of risk when the costs of type II errors are changed compared to the *L* = 9% level.

Clearly, then, the VaR at a 99% level of reliability is considerably more sensitive to input parameter *D*. Where *D* is increased by 1% – ceteris paribus – the value of the VaR will increase by 0.93%, while the change is a mere 0.03% where *L* is increased. It is also clear that the extent of change is not symmetrical. For example, with an increase in the costs of both types of error at the same time, the VaR will increase by 1.02%; whereas if both are reduced, the VaR will decrease by just 0.91%. A credit institution must bear these findings in mind when estimating input parameters *D* and *L* in measuring model risk, because the VaR is sensitive to them.

To round off this analysis, it is also worth asking whether the extents of risk quantified above are in fact an accurate measure of model risk. Naturally the answer is "no," since risk is a latent concept, meaning that it cannot be measured directly. As such, "the value of any quantifiable will always be an approximation, highlighting a single factor of theoretically complex, multi-layer and multi-factor real risk" (*Bélyácz*, 2011, p. 309).

5. SUMMARY

This paper set out to measure model risk in credit scoring models, with a view to providing credit institutions a better understanding of possible losses resulting from the models they use, and using the information to support executives in making decisions.

The definition of model risk was followed by a brief survey of relevant legislation within the Basel framework. This was followed by an examination of a minor, but all the more crucial, area of the problem: model risk in credit scoring models.

Next, the paper dealt with the question of the representativity of basic data. The problem known as selection bias occurs because banks only possess a complete data set for applicants who have already been granted a loan. For clients whom the credit institution has rejected, the changing value of credit risk is not known; that is, what would have happened if they had been granted a loan. Consequently, model risk can be regarded as a problem of missing data.

Bearing this in mind, we reviewed the main types of missing data, focusing on a method based on multiple imputation. The reason we elected to use this particular method from among several available options is because it affords inclusion in the estimation of the insecurity caused by missing data, which is crucial with respect to risk management.

Next, the paper showed how, for given probabilities of default and the default variable, the cutoff can be used to model confusion matrices of different types, and in turn, to calculate model risk losses. Applying the model of threshold exceedances to the tails of loss distribution, the different values of model risk can easily be determined.

After presenting the theoretical basics, we proceeded to a demonstration using a practical example. Following a brief description of the data set, we applied a multiple imputation procedure, assuming single-variable data missing at random (MAR), to calculate the *default* variable of rejected clients. It was found that the bootstrap approach led to more accurate estimations.

Next, we estimated the probability of default for individual clients, on the basis of which we calculated possible model risk-related losses. Applying the generalized Pareto distribution to the tails of loss distribution, we determined two risk measures: the VaR and expected shortfall.

Subsequently we examined the sensitivity of the VaR at a 99% level of reliability, first to the threshold used in the application of the extreme value theory, and secondly to the costs of type I and II errors. It was concluded that the VaR is greatly reliable in choosing the threshold; however, the values of individual losses resulting from model errors significantly affect the risk measure, making their most accurate estimation of vital importance to the final outcome.

As the title of this paper points out, a perfect model cannot, by definition, exist; in other words, comprehensively charting model error-related dangers is an essential task. The procedure used by the authors of this paper likewise presents only a simplified section of the infinitely complex area of model risk. Recalling the words of *George E. P. Box*, these efforts cannot claim to produce faultless results either, but we nevertheless hope that they may prove valuable in identifying model risk and making relevant banking decisions.

REFERENCES

Bélyácz, I. (2011): Risk, uncertainty, probability. In: *Hitelintézeti Szemle* , Vol. 10, No. 4, pp. 289–313.

Box, G. E. P. – Draper, N. R. (2007): *Response Surfaces, Mixtures, and Ridge Analyses*. Second Edition, New Jersey: John Wiley & Sons.

Buuren, S. – Groothuis-Oudshoorn, K. (2011): Multivariate Imputation by Chained Equations. In: *R. Journal of Statistical Software*, Vol. 45, No. 3, pp. 1–67.

Gilleland, E. – Ribatet, M. – Stephenson, A. G. (2013): A software review for extreme value analysis. *Extremes* 16, pp. 103–119.

Hofmann, H. (1994): German Credit Data – data set. Downloadable from: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) Date of download: 9 October 2014.

Jorion, P. (1999): *Value at Risk*. Budapest: Panem Kiadó.

Kiss, F. (2003): *The development and application of credit scoring*. PhD thesis, Budapest University of Technology and Economics.

Kovács, E. (2011): *A statistical analysis of financial data*. Tanszék Kiadó, Budapest.

Krekó, B. (2011): Risk, uncertainty and model risk from a risk management perspective. In: *Hitelintézeti Szemle*, Vol. 10, No. 4, pp. 370–378.

Little, R. J. A. – Rubin, D. B. (2002): *Statistical Analysis with Missing Data*. Second Edition, New Jersey: John Wiley & Sons.

McNeil, A. J. – Frey, R. – Embrechts, P. (2005): *Quantitative Risk Management. Concepts, Techniques and Tools*. New Jersey: Princeton University Press.

McNeil, A. J. (1999): *Extreme Value Theory for Risk Managers*. London: Risk Books, 1999, pp. 93–113.

National Bank of Hungary (2012): The Supervisory Review Process (SRP). A methodological guide. May 2012.

Oravecz, B. (2008): *Selection bias and its reduction in credit scoring models*. PhD thesis, Corvinus University of Budapest, Doctoral School of Business and Management.

Radnai, M. – Vonnák, D. (2010): *Bank capital adequacy manual*. Budapest: Alinea Kiadó.

Regulation (EU) No. 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No. 648/2012.

Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Thomas, L. C. – Edelman, D. B. – Crook, J. N. (2002): *Credit Scoring and Its Applications*. Philadelphia: Society for Industrial and Applied Mathematics.

Tulassay, Zs. (2013): *Extreme Value Theory*. In: Daróczi, G. – Puhle, M. – Berlinger, E. – Csóka, P. – Havran, D. – Michaletzky, M. – Tulassay, Zs. – Váradi, K. – Vidovics-Dancs, Á. (eds.): *Introduction to R for Quantitative Finance*. Birmingham: Packt Publishing.

White, I. R. – Daniel, R. – Royston, P. (2010): Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis*, Vol. 54, Issue 10, pp. 2267–2275.