BANCA D'ITALIA
EUROSISTEMA

Temi di discussione

del Servizio Studi

A neural network architecture for data editing in the Bank of Italy's business surveys

by Claudia Biancotti, Leandro D'Aurizio and Raffaele Tartaglia Polcini

Number 612 - February 2007

The purpose of the Temi di discussione *series is to promote the circulation of working papers prepared within the Bank of Italy or presented in Bank seminars by outside economists with the aim of stimulating comments and suggestions.*

The views expressed in the articles are those of the authors and do not involve the responsibility of the Bank.

A neural network architecture for data editing in the Bank of Italy's business surveys

Claudia Biancotti(*), Leandro D'Aurizio(*) and Raffaele Tartaglia Polcini(*)

Abstract

This paper presents an application of neural network models to predictive classification for data quality control. Our aim is to identify data affected by measurement error in the Bank of Italy's business surveys. We build an architecture consisting of three feed-forward networks for variables related to employment, sales and investment respectively: the networks are trained on input matrices extracted from the error-free final survey database for the 2003 wave, and subjected to stochastic transformations reproducing known error patterns. A binary indicator of unit perturbation is used as the output variable. The networks are trained with the Resilient Propagation learning algorithm. On the training and validation sets, correct predictions occur in about 90 per cent of the records for employment, 94 per cent for sales, and 75 per cent for investment. On independent test sets, the respective quotas average 92, 80 and 70 per cent. On our data, neural networks perform much better as classifiers than logistic regression, one of the most popular competing methods. They appear to provide a valid means of improving the efficiency of the quality control process and, ultimately, the reliability of survey data.

Keywords: data quality, data editing, binary classification, neural networks, measurement error.
JEL codes: C45, C42.

CONTENTS

(*) Bank of Italy, Economic Research Department.

1. Introduction

Attention to quality has become a central preoccupation for data producers. In a statistical framework, quality control is routinely called "editing", and is widely defined as follows: “[data] *editing is the activity aimed at gathering intelligence related to significant differences in the data for analytical purposes, providing feedback that can lead to improvements in data collection and processing, reducing the level of error present in the data and ensuring a degree of consistency, integrity, and coherence*” (Chinnappa et al., 1990). Under the Total Quality Management approach, quality equates to fitness of survey microdata for purposes of research or statistical information. Brackstone (1999) lists six dimensions of quality: *relevance, accuracy, timeliness, accessibility, interpretability* and *coherence.* The concept of interpretability is also meant to encompass the provision of useful indications about the accuracy of information, in the guise of documentation and metadata. In this context, interpretability is enhanced both by making the user aware of the correct application of the editing methods employed to rectify wrong data, and by providing synthetical indicators of the anomalies detected. Granquist and Kovar (1997) give a thorough account of the resource problems stemming from the need for data quality control, and relate some problems associated with the traditional editing methods. According to them, quality-related costs may amount to as much as 40 per cent of the whole cost of a typical business survey. Open issues relating to data quality control are discussed, for example, in Rivière (2002).

Strategies for automated and semi-automated error detection are constantly improving. Macro-level editing methods are mostly concerned with maximizing the reliability of summary statistics and regression-based estimates. In an effort towards timeliness and cost control, measurement error resulting in serious distortions of the final results is tackled, while mistakes of little consequence are overlooked (see for example Battipaglia, 2002). A macro-editing procedure normally ensures quality enhancement of the final estimates by selecting significant units to be re-contacted. The most evident drawback of these techniques is their low “hit rate” (i.e. the share of the number of flags that result in changes of the original data; Hedlin, 2003): for re-contacted firms, confirmations of apparently anomalous values are far more numerous than revisions. There is also a price to pay: progressive segmentation of the sample between frequently monitored units (since they are more influential on the final estimate) and less frequently monitored (less influential) units occurs. Typically, for estimates of totals and rates of change, the big firms are more likely to be considered influential and therefore re-contacted.

Micro-level approaches, generally more time-consuming and expensive, complement the macro approach and may provide a remedy to this problem by allowing a reliability assessment of single

items, regardless of their impact on a pre-defined set of estimates. This is very important whenever data end up being used for a wide variety of research purposes that are not predictable *ex ante*.

This paper explores the use of artificial neural networks (ANNs) as a micro-editing tool for the Bank of Italy's business surveys. It has a strong applied orientation, dealing with a specific neural network architecture selected for error detection and data editing for the surveys. According to Breiman (1994, 2001), the very nature of neural network modelling makes it highly dependent on the particular data at hand and the search of the best network in terms of predictive accuracy is strongly problem-oriented. Section 2 briefly makes the case for the use of neural networks in our context. Section 3 presents the surveys and their data quality issues. Section 4 provides details of the data sets used for our experiment. Section 5 discusses network architecture, estimation strategies and results. Section 6 concludes and indicates the way forward for further developments. For the sake of brevity, the methodological tenets of neural computing are not explained in the paper: the Appendix provides a brief summary and references.

2. The case for neural networks

Supporters and critics of ANNs balance their arguments. A proven high predictive capability stands against some unresolved issues that still punctuate the existing theoretical framework: for example, the absence of optimality criteria for the choice of the network topology (e.g. how many hidden layers, how many hidden units, how to place arcs), the activation function, the learning algorithm, the weight adjustment rule; the want of a quality assessment for the predicted values.

From an applied mathematical perspective, however, the developers of neural networks can claim the *universal approximation property* (Hornik, Stinchcombe and White, 1989; for a heuristic, suggestive proof and extensions, see Ripley, 1996): any real-valued, continuous function of real variables can be uniformly approximated on compacta by a neural network with a logistic or threshold squasher and one hidden layer.

From this point of view, ANNs are computer-intensive, specification-free nonlinear fitters; many existing multivariate regression models can be seen as restrictions of this general framework. In our research, aimed at efficient error-spotting and effective data quality enhancement rather than interpretability, ANNs are a tool allowing detection of mistakes on many different variables, taking into account their interaction without having to describe the connections between observed features of the respondent and erring propensity by way of possibly under-performing "structural" equations.

Neural networks have entered the economist's toolbox fairly recently. An overview of existing applications is found, for example, in Graepel et al. (2001). Although neural computing is also a relatively new tool in editing, studies on the subject can be traced back to Roddick (1995);

discussions from a mixed theoretical-empirical standpoint can be found in Nordbotten (1995, 1996), Larsen and Madsen (1999), Biancotti and Tartaglia-Polcini (2005) among others. Implementation details of ANN-based editing strategies feature in several operational manuals of national statistical offices and research institutions. Machine-learning methods for predictive classification are interesting in that they have the potential to outperform their traditional counterparts (such as discriminant analysis or logistic regression) because of their ability to adapt quickly to changes in the structure of phenomena (Rohwer, Wynne-Jones and Wysotzki, 1997). In laymen's terms, if we can train a computer to see mistakes in a dataset, without being too explicit about what kind of error to look for, we can save vast amounts of time and achieve greater accuracy under given budget and time constraints.

3. Data quality in the Bank of Italy's business surveys

The Bank of Italy has been carrying out business surveys on a yearly basis since 1972. Until 1998 only manufacturing firms with more than 49 employees were included in the sample; the reference population gradually grew from the following year, and now covers firms with at least 20 employees belonging to both the industrial sector and private services (excluding banking and insurance). The sample includes some 4,000 firms. Units were originally chosen at random under a stratified design and always re-contacted in all waves following the one during which they had been drawn, provided that they still belonged to the target population. Refusals and firms no longer in the target population are routinely replaced with similar units. Interviews are made in the first months of the year *t+1* and concern data about the years *t-1* and *t*, together with forecasts for the current year *t+1*. Data collection is conducted within the Bank, through its local branches.

The variables collected range from levels of sales and investment to indebtedness and other sources of financing (expressed in thousands of euro). Typical sources of measurement error are:

1. response in euros instead of thousands of euros, or in former national currency units instead of euros: this problem was particularly acute in the years immediately following the adoption of the new currency;
2. misclassification of aggregates to be included in the amounts;
3. mergers or acquisitions not correctly handled;[2]

[2] Firms that have been through mergers or acquisitions are considered only if data come from the same set of local units and employees for the three years *t-1*, *t* and *t+1*. This is achieved either by fictitiously pre-dating the merger/acquisition event to the beginning of year *t-1* or by postponing it to the end of year *t+1*. Measurement error arises from mistakes in these adjustments.

4. misreadings in the paper questionnaire (either manual or through optical character recognition).

Data quality is taken care of in successive steps. Simple checks are already implemented at the data-entry level with the use of CAPI (Computer Aided Personal Interviewing).[4] A further level of control consists in checking for admissible ranges of quantitative variables: based on past distributions or specialists' opinions, values outside the expected range must be double-checked by the interviewer and a flag activated if the value is confirmed. The final stage relies on a selective editing procedure. A linear model for the published estimates (for example, rates of change for sales and investment) is fitted and the predicted value is trusted as the "true" one under the model. A first-order Taylor approximation (the "score") evaluates the contribution of the single units to this prediction. Units with the highest score are labelled anomalous; these are normally re-contacted.

4. The data

4.1 The initial dataset and the error-generating process

In order to conduct our experiment we build a dataset based on actual business survey microdata. We generate errors by perturbing large subsets of correctly valued quantitative survey variables, which are more prone to measurement errors.

We use simulated errors instead of historically recorded ones because by doing so we can accurately keep track of correct and wrong data.

We need to produce a significant per centage of errors, since the estimation of a neural network for classification calls for a fair-sized class of wrong records. The choice enables the training to yield robust results, given that training an artificial neural network for classification is an iterative process liable to produce wrong results if one class size is too low. The frequency of correct data hovers around the threshold of 50-55 per cent and is certainly too low if compared with real survey scenarios; however, as discussed below, the set of rules upon which the perturbations are based carefully reproduce actual error patterns to provide a realistic training set for the network. The learning process is not biased by the high frequency of errors, which is merely instrumental to its iterative nature.

Our simulated datasets are based on the final archives for the business surveys conducted during 2004. Typically, questions are asked about the two previous years (2002 and 2003 in this case). We

[4] For example, negative values are not accepted if they do not have economic meaning; answers to discretely coded variables can only belong to a pre-defined list.

assume that these archives include no leftover errors, having undergone the control and editing procedures detailed in Section 3.

Five main categories of perturbations are introduced, covering the widest possible range of error types: substitution of the original value of a variable with a zero; generation or elimination of a zero digit in a variable; units/thousands/millions arbitrary transformations (addition or subtraction of three zero digits at the end of a variable); swapping of values of a randomly chosen variable between two firms surveyed by the same interviewer; random generation of tail values[5] from the empirical distribution, in order to account for residual sources of measurement error.

4.2 Selection of the variables

The selection of the variables to be considered in the experiment poses a number of problems. The scope of our networks should not be confined to predicting whether a particular record contains an error somewhere. While already a relatively interesting accomplishment from the machine-learning standpoint, it is not operationally useful since a generic indication would force us to go through all the suspect variables manually in order to single out the mistake. We need some pointers to where the error is within the record, ideally to the level of the specific variable. The perfect editing device should be trained on all the 200-odd survey questions, but this is easier said than done. A first important hurdle is our estimation algorithm, which – in the fashion of nearly all estimation algorithms – does not allow for missing values: we are forced to exclude variables that are seriously affected by item non-response. The elimination of all records with at least one missing element would train the network on a sample that is too narrow; performing estimates on the basis of imputed values would be even worse because standard methods of imputation artificially reduce the variance in the regressors, damaging the informational content of the results. Moreover, in order to obtain results that can be extended to a typical wave of the survey we have to exclude one-shot sections: alternatively, the model should be modified for every wave of the survey. We also forgo questions routinely asked only to a part of the sample, because they might have peculiar error patterns, correlated with the very inclusion of a firm in the subsample. Forecasts are left out because of the impossibility of separating measurement error from forecasting error, and of the high frequency of missing values.

These difficulties lead us to choose the solution of disclosing errors only for variables related to the three core topics investigated in the questionnaire: employment, sales and investment. These phenomena, anyway, deserve maximum priority in any plan of quality improvement: the survey's main objective is to evaluate short-term dynamics of macroeconomic aggregates. The relevant

[5] These perturbed data are obtained by the following steps: the value to perturb is placed in the lower or upper tail if it is respectively below or above the median. A random per centile is then extracted to define the tail dimension and a final random drawing sets the position of the perturbed value inside the tail.

estimates are the only ones that get published every year without exception, and they tend to draw the most attention from the readership.

There are twenty such core variables: twelve for employment, four for sales, and four for investment. The perturbation algorithms presented in Section 5 are applied to these twenty variables, and then three neural networks are separately trained to localize the general presence of errors within each of the groups. We made this choice because the algorithm did not work when it tried to identify errors at the single variable level. A balance was struck between desired resolution power and performance by training the neural network to localize the general presence of error within each of the groups: this choice also allows us to account for the relationships among the different variables within the same group.

Some non-perturbed stratification variables, such as the geographical location and the sector of activity of the firm, are also kept on the restricted datasets, which comprise 2,959 observations for employment, 2,968 for sales and 2,962 for investment, out of around 4,200 firms participating in the survey.

4.3 The training and validation datasets

In order to train the networks, we produce three datasets of the same size with generated errors for employment, sales and investment. Mistakes are generated independently for the three categories of variables, each containing 45-50 per cent of wrong data. Each dataset is then randomly split into two equally sized samples, to be used respectively for training and validation.

Network architecture building starts from a decision on how to submit the data to the network: the same set of information can be presented in a variety of ways that differ considerably in explanatory power.

Since the chosen neural network can be thought of as an enhancement of a multinomial logit model, we assumed that the use of stratification variables (strictly related to sampling design) as inputs would improve the classification task. This is proven to be effective only for sales, whereas we resolved to get rid of them for employment and investment, as they showed no contribution to learning.

As a rule, raw variables possess a limited solving potential and must be flanked by some transforms, typically ratios. Ratios works better than the raw input variables because a neural network, as much as any other binary classifier, has the primary goal of dividing space into acceptance and rejection regions. This will be harder as the dimensionality of the space increases, especially when the type of partition that has to be found in each dimension is not trivial. This is particularly true for outlier detection: the presence of “low” and “high” outliers would call for a partition of the space in separate regions. No banal threshold function would work; if we had used hidden nodes with a logistic activation, we would have needed at least two nodes to perform this

very simple activity on a single variable.[7] Ratios reduce the dimensionality of the problem, allowing for lighter networks and reducing the risk of overfitting, and are, moreover, able to turn data consistency problems into simpler range problems.

This point can be highlighted through a simple example: suppose the national firm-level average of yearly hours worked is around 300,000. The network must be trained to understand that both 1,000 and 5,000,000 are wrong, but must also beware if a firm declares 150,000 in 2002 and 500,000 in 2003; if only the raw values are used as inputs, a very large number of hidden nodes is needed to get a decent performance. If the ratio of hours worked per employee in 2002 to hours worked per employee in 2003 is used as input, the problem is reduced to finding ratios outside a reasonable range (say 0.8 to 1.2). Cases of both values multiplied by 1,000 are not caught, but they are anyway marked as wrong by adding a specific control on yearly hours worked per capita.

5. Network architecture, estimation strategy and results

5.1 The general features of the architecture

Three separate networks are trained: one catches mistakes in employment variables, one in sales, one in investment. We find that errors in employment can be singled out based on predictors of the same group, e.g. number of hours worked per employee, or per centage of overtime. In the case of sales and investment, the task is not quite as self-contained and some employment-related predictors are also needed, in the form of ratios per employee. This result calls for some extra care in how the networks are used: while a wrong number on investment will not affect how the system detects mistakes on employment, the opposite is not true. An operational hierarchy is necessary: the employment network is to be run first and then the other networks, after getting rid of the mistakes singled out by the first one.

Topologically, the networks are built on the traditional feed-forward scheme. They feature large hidden layers with logistic activations, with the addition of one or two discrete-value nodes, directly connected with input nodes. Output nodes are also logistic, but the result is clipped to a dummy variable indicating the erroneous state of the observation depending on the original output value (see the Appendix).

As for the estimation strategy, we always use the Resilient Propagation (Rprop) algorithm: traditional learning devices such as vanilla back-propagation, or slight variations of it, were no

[7] If we worked on two variables, at least four hidden nodes would be required to spot both “low” and “high” outliers in each of those. One experiment showed that if both variables took wrong albeit reasonable values, the network could not extrapolate a rule, because of a dominating range effect.

match for our problems and did not produce good performances in terms of mean square error (MSE). The mechanics of the Rprop algorithm are quite simple: signs are used instead of levels ("Manhattan learning") for derivatives and the learning rate is dynamically updated (Riedmiller and Braun 1993; see the Appendix). This has been shown to be suitable for data as heavy-tailed as ours.

The hidden nodes can be seen as detectors of abstract features of the space of variables, so that their number accounts for the dimension of the (possibly non-linear) principal components' space that filters the lower-variance noise signal (Rohwer, Wynne-Jones and Wysotzki, 1994). We pruned the hidden nodes, based on the absolute magnitude of the weights (see the Appendix), keeping an eye simultaneously on training and validation error, so as to avoid overfitting and preserve parsimony. We also tried more than one hidden layer, with no better results: this is not surprising, given the universal approximation property enjoyed by feed-forward neural networks with one hidden layer.

5.2 The evaluation

The random perturbation of data regarded as "correct" and the following analysis of how well "wrong" data are pinpointed is a classical tool to evaluate editing techniques (ET) (see, for example, Barcaroli and D'Aurizio, 1997). Simple indexes help us in the assessment, for example the Error Identification Capability (EIC), defined as the per centage of correctly identified wrong values. Two kinds of errors feature in any editing process: a) marking correct data as erroneous (identification error of wrong values); b) marking errors as correct data (identification error of correct values). An ideal ET should keep the frequency of the two errors to a minimum.

These simple evaluation tools are illustrated by the following two-way contingency table, which we will use in the course of our analysis.

True value	**Predicted value**	
	correct	wrong
correct	N_{11}	N_{12}
wrong	N_{21}	N_{22}

The EIC is $\frac{N_{22}}{N_{12}+N_{22}}*100$, while the two previously mentioned errors are the cases corresponding to the cells containing the values N_{12} and N_{21}: they are kept to a minimum if $N_{11}+ N_{22}$ (number of correct predictions) is as close as possible to the total $N_{11}+ N_{12}+ N_{21}+ N_{22}$.

For each experiment, the network performance will be assessed against the logistic model[8] (a traditional binary classifier) for the training set, for the validation set and for two test sets: the latter are two equally sized samples drawn from the original dataset and comprising all types of perturbations. Each table cell presents the number of cases and the total, row and column per centages.

In order to evaluate the network stability, we also present a graph with the MSEs on the training and validation datasets against the number of iterations (the MSEs for the test sets are omitted for brevity). The two plots should ideally decrease and keep close after the first iterations.

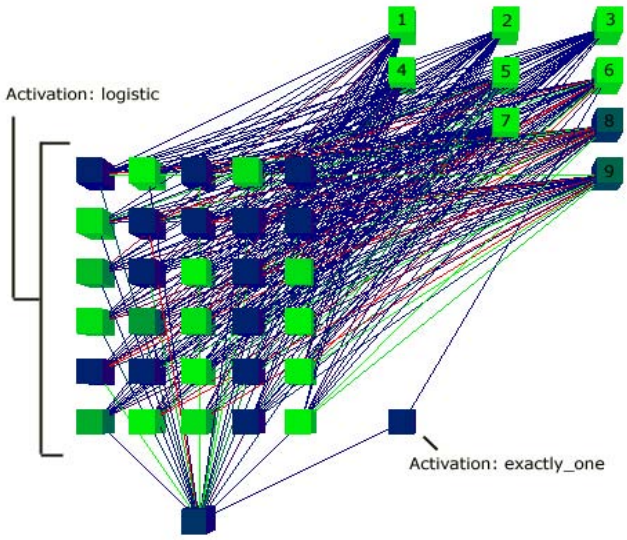
5.3 Employment

The employment variables affected by error in our dataset are as follows: average number of employees, end-of- year number of employees, total number of hours worked in a year, per centage of overtime hours, number of new hirings, number of departures. Values for the year 2002 and the year 2003 (both drawn from the 2004 wave of the survey) are provided for each of those variables. As stated above, raw variables showed a rather limited predictive power: after an extensive number of trials, nine transforms were selected for inclusion in the network. Figure 1 illustrates the network topology; the list of variables is in the legend below. The learning and validation curves are shown in Figure 2. The performance of the network is detailed in Tables 1a-1d.

[8] The logistic models presented feature the same input variables used for the neural network. We also tried to use other covariates, including all the stratification variables included in the survey design, but with no substantial improvement in the results.

Figure 1

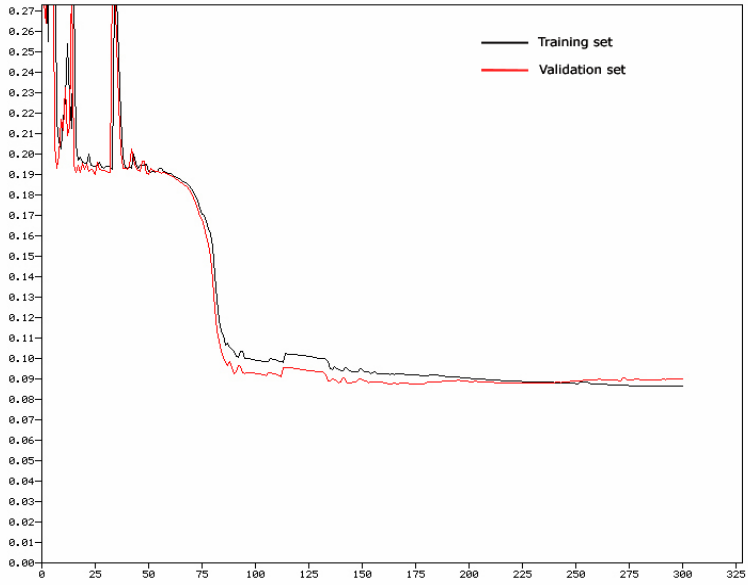
Neural network for error detection: Employment



1. Ratio of total hours worked, 2002, to total hours worked, 2003; 2. Hours worked per employee, 2002; 3. Hours worked per employee, 2003; 4. Ratio of hours worked per employee, 2002, to hours worked per employee, 2003; 5. Ratio of average number of employees, 2002, to average number of employees, 2003; 6. Consistency check: takes the value 1 when the end-of-year number of employees, 2003, corresponds to the sum of the end-of-year number of employees, 2002, and the new hirings during 2003 minus the departures during 2003, and 0 otherwise; 7. Ratio of share of overtime, 2002, to share of overtime, 2003; 8. Ratio of new hirings in 2002 to average number of employees, 2002; 9. Ratio of departures in 2002 to average number of employees, 2002.

Figure 2

Training and validation MSE: Employment



Tables 1a-1d

EMPLOYMENT: CONFUSION MATRIX FOR THE TRAINING SET
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	796 53.78 100.00 84.14	791 53.45 99.37 73.72	0 0.00 0.00 0.00	5 0.34 0.63 1.23	796 53.78
	Wrong	150 10.14 21.93 15.86	282 19.05 41.23 26.28	534 36.08 78.07 100.00	402 27.16 58.77 98.77	684 46.22
	Total	946 63.92	1073 72.50	534 36.08	407 27.50	1480 100.00

Correct predictions: ANN 89.86 per cent; logistic regression 80.61 per cent

EMPLOYMENT: CONFUSION MATRIX FOR THE VALIDATION SET
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	817 55.24 98.55 85.64	822 55.58 99.16 76.32	12 0.81 1.45 2.29	7 0.47 0.84 1.74	829 56.05
	Wrong	137 9.26 21.08 14.36	255 17.24 39.23 23.68	513 34.69 78.92 97.71	395 26.71 60.77 98.26	650 43.95
	Total	954 64.5	1077 72.82	525 35.5	402 27.18	1479 100.00

Correct predictions: ANN 89.93 per cent; logistic regression 82.29 per cent

EMPLOYMENT: CONFUSION MATRIX FOR THE TEST SET I
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	1109 74.98 99.02 90.75	1107 74.85 98.84 86.48	11 0.74 0.98 4.28	13 0.88 1.16 6.53	1120 75.73
	Wrong	113 7.64 31.48 9.25	173 11.7 48.19 13.52	246 16.63 68.52 95.72	186 12.58 51.81 93.47	359 24.27
	Total	1222 82.62	1280 86.54	257 17.38	199 13.46	1479 100.00

Correct predictions: ANN 91.61 per cent; logistic regression 87.43 per cent

EMPLOYMENT: CONFUSION MATRIX FOR THE TEST SET II
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	1114 75.27 99.02 91.09	1115 75.34 99.11 86.37	11 0.74 0.98 4.28	10 0.68 0.89 5.29	1125 76.01
	Wrong	109 7.36 30.7 8.91	176 11.89 49.58 13.63	246 16.62 69.3 95.72	179 12.09 50.42 94.71	355 23.99
	Total	1223 82.64	1291 87.23	257 17.36	189 12.77	1480 100.00

Correct predictions: ANN 91.89 per cent; logistic regression 85.43 per cent

On the training dataset, 100 per cent of non-erroneous records are recognized correctly; conditional on an error being present, the hit rate is 78.07 per cent. The network shows a satisfactory performance: the logistic model, totals 99.37 per cent on correct information, but only 58.77 per cent on erroneous records. By looking at individual data, we find that the mistakes not seen by the network are, typically, those that are neither outliers nor are able to alter the ratio structure. A perturbed value is not spotted if it falls within a reasonable range and its use in subsequent estimations does not produce anything weird, meaning that the network is as bad at understanding mistakes resembling correct information as human experts may be.

The generalization shows that the results change very little when the network is fitted on the validation dataset: 99.82 per cent of the non-erroneous records and 78.92 of the erroneous ones are correctly classified. On the test sets, the hit rate is still above 99 per cent for the non-perturbed records, but falls to around 70 per cent for the ones containing mistakes. Microdata inspection shows that this behaviour mostly depends on unspotted digit-swapping perturbations, whenever they do not affect the first digits on the left.

5.4 Sales

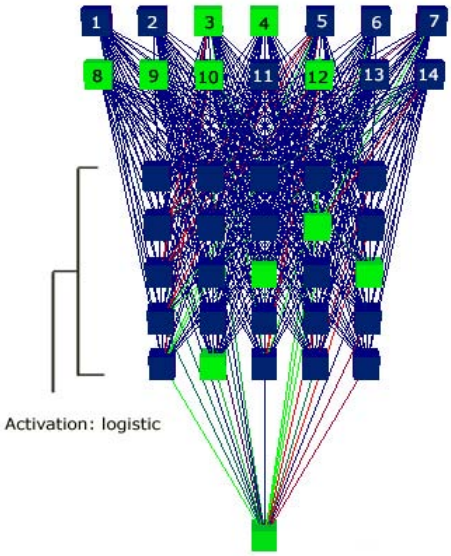
Four variables concerning the firm's overall sales are included in the simulation experiment: total sales for 2002 and 2003, together with the export sales for both years. The implemented network draws largely on the experience accumulated in setting up the one for the employment variables: transforms of the original variables are similarly included, together with the original levels that are to be checked.

The two ratios introduced are the level of total sales per employee and the quota of export sales over total sales. In this case, too, ratios are useful alongside raw variables not so much as a shield against risks of overfitting (less likely than in the case of employment due to the smaller number of variables involved) but rather for the fundamental role they play in speeding up the learning and generalization performances of the network.

Dichotomous variables indicating the presence of zeroes are added: they are a sure indication of error for total sales (a firm can hardly survive without making any sales over a whole year). For export sales these indicators, together with the information about firm size and sector of economic activity, can help pinpoint suspicious zeroes. Inspection of micro data shows that the stratification variables explain more variability for sales than employment: particularly for export sales, the combined information of economic activity and firm size acts as a powerful discriminant in separating "true" from "false" zeros. The topology is shown in Figure 3. Figure 4 shows the MSE evaluated on the training set and the validation set. Tables 2a – 2d present the performance of the network.

Figure 3

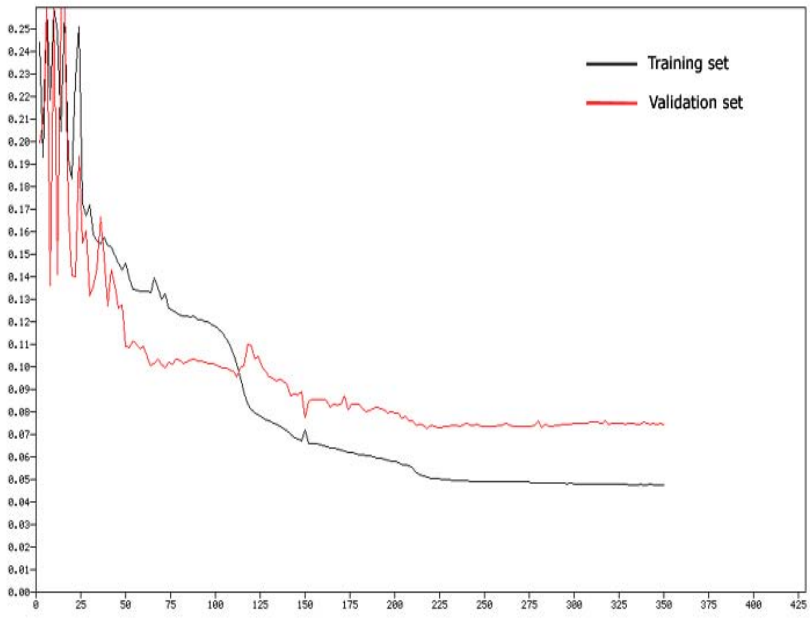
Neural network for error detection: Sales



1. Firm size class; 2. Sector of economic activity; 3. Total sales, 2002; 4. Total sales, 2003; 5. Export sales, 2002; 6. Export sales, 2003; 7. Total sales per employee, 2002; 8. Total sales per employee, 2003; 9. Export sales per employee, 2002; 10. Export sales per employee, 2003; 11. Boolean indicator taking the value 1 if total sales, 2002 = 0, and 0 otherwise; 12. Boolean indicator taking the value 1 if total sales, 2003 = 0, and 0 otherwise; 13. Boolean indicator taking the value 1 if export sales, 2002 = 0, and 0 otherwise; 14. Boolean indicator taking the value 1 if export sales, 2003 = 0, and 0 otherwise.

Figure 4

Training and validation MSE: Sales



Tables 2a-2d

SALES: CONFUSION MATRIX FOR THE TRAINING SET

(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	702 47.3 96.43 94.99	621 41.85 85.3 63.37	26 1.75 3.57 3.49	107 7.21 14.7 21.23	728 49.06
	Wrong	37 2.49 4.89 5.01	359 24.19 47.49 36.63	719 48.45 95.11 96.51	397 26.75 52.51 78.77	756 50.94
	Total	739 49.8	980 66.04	745 50.2	504 33.96	1484 100.00

Correct predictions: ANN 95.75 per cent; logistic regression 68.60 per cent

SALES: CONFUSION MATRIX FOR THE VALIDATION SET

(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	875 58.96 95.21 93.88	581 39.15 63.22 68.03	44 2.96 4.79 7.97	338 22.78 36.78 53.65	919 61.93
	Wrong	57 3.84 10.09 6.12	273 18.4 48.32 31.97	508 34.23 89.91 92.03	292 19.68 51.68 46.35	565 38.07
	Total	739 49.8	980 66.04	745 50.2	504 33.96	1484 100.00

Correct predictions: ANN 93.19 per cent; logistic regression 58.83 per cent

SALES: CONFUSION MATRIX FOR THE TEST SET I

(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	622 41.91 84.74 80.88	638 42.99 86.92 62.24	112 7.55 15.26 15.66	96 6.47 13.08 20.92	734 49.46
	Wrong	147 9.91 19.6 19.12	387 26.08 51.6 37.76	603 40.63 80.4 84.34	363 24.46 48.4 79.08	750 50.54
	Total	769 51.82	1025 69.07	715 48.18	459 30.93	1484 100.00

Correct predictions: ANN 82.54 per cent; logistic regression 67.45 per cent

SALES: CONFUSION MATRIX FOR THE TEST SET II

(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	616 41.51 68.44 87.62	618 41.64 68.67 68.82	284 19.14 31.56 36.36	282 19 31.33 48.12	900 60.65
	Wrong	87 5.86 14.9 12.38	280 18.87 47.95 31.18	497 33.49 85.1 63.64	304 20.49 52.05 51.88	584 39.35
	Total	703 47.37	898 60.51	781 52.63	586 39.49	1484 100.00

Correct predictions: ANN 75.00 per cent; logistic regression 62.13 per cent.

The training produces an almost perfect match between true and predicted error flags (95.75 per cent of matches and 95.11 per cent of true errors correctly identified). The same good performance is shown for the validation sample. Results are obviously less good, but still acceptable, when the network is run on the test sets. The logistic model is again outperformed.

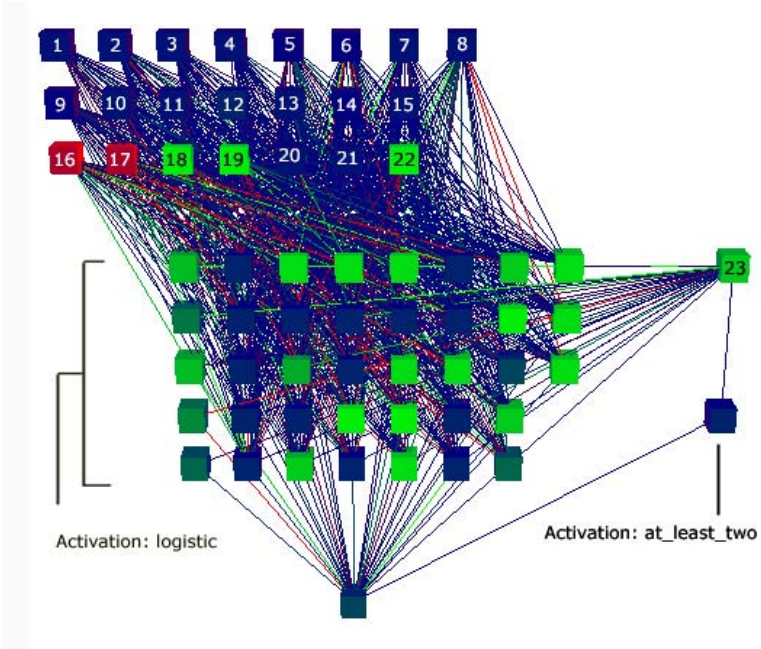
5.5 Investment

There are only four raw survey variables concerning investment: equipment, machinery and real estate in 2002; software and intellectual property in 2002; the same two variables, but for 2003. In order to train the network properly, however, we need more variables. Figure 5 shows the network topology and presents a short description of the 23 input variables. These can be divided into four broad groups: a) levels of the original variables; b) ratios calculated with respect to non-investment variables (e.g. investment per employee, investment to sales); c) variations in levels and ratios; d) dummy indicators of data structure. The number of employees needed to build some of the ratios are taken from the non-perturbed dataset: in assuming they are correct, we are embracing the hierarchical approach advocated in Section 5.1.

After a first glance at the list of inputs, one can already tell that predicting errors in investment is noticeably more difficult than catching mistakes in employment or sales. The intrinsic variability of the phenomenon is much larger and much less patterned: for example, it is possible for a firm to invest a huge amount in a given year and very little in the following year. To make matters worse, two firms can be very similar in other respects and have substantially different investment behaviour: it then becomes almost impossible to find a contrast term in order to tell which part of the variability comes from an error and which comes from real-life heterogeneity. Finally, a special problem emerges where the treatment of zeros is concerned. Some of them are actually true: it is plausible that a firm stops investing for one year. Some are false and should be substituted by some unknown amount, others are equally false, but should be classified instead as item non-response.

Figure 5

Neural network for error detection: Investment



1. Tangible investment, 2002; 2. Tangible investment, 2003; 3. Intangible investment, 2002; 4. Intangible investment, 2003; 5. Tangible Investment per employee, 2002; 6 Tangible Investment per employee, 2003; 7. Intangible Investment per employee, 2002; 8. Intangible Investment per employee, 2003; 9. Ratio of Tangible investment to total sales, 2002; 10.Ratio of Tangible investment to total sales, 2003; 11. Ratio of Intangible investment to total sales, 2002; 12. Ratio of Intangible investment to total sales, 2003; 13. Sum of total investment, 2002, and total investment, 2003; 14. Total investment per employee, calculated based on (13) and average employment, 2002 and 2003; 15. Tangible investment: signed variation between 2002 and 2003; 16. Intangible investment: signed variation between 2002 and 2003; 17. Tangible investment: signed relative variation between 2002 and 2003; 18. Intangible investment: signed relative variation between 2002 and 2003; 19. Boolean indicator taking the value 1 if Tangible investment, 2002 = 0, and 0 otherwise; 20. Boolean indicator taking the value 1 if tangible investment, 2003 = 0, and 0 otherwise; 21. Boolean indicator taking the value 1 if intangible investment, 2002 = 0, and 0 otherwise; 22. Boolean indicator taking the value 1 if intangible investment, 2003 = 0, and 0 otherwise; 23. Sum of the four boolean indicators sub (19), (20), (21), and (22.).

This scenario implies that the approach used in the case of employment, i.e. a straightforward reduction of the problem dimensionality by moulding it into a combination of several simple range problems and consistency checks, cannot work for investment. The concept of acceptable range is less clearly defined: if a structure similar to the one *sub* 5.3 were proposed, only the few major outliers would be caught. Prior knowledge of a different kind is needed by the network. Through trial and error we find that the level, ratio and variation variables are all useful to fine-tune the performance of the algorithm, but the big jump in performance is achieved by introducing a set of four dummy variables that detail whether each of the four raw variables is zero-valued.

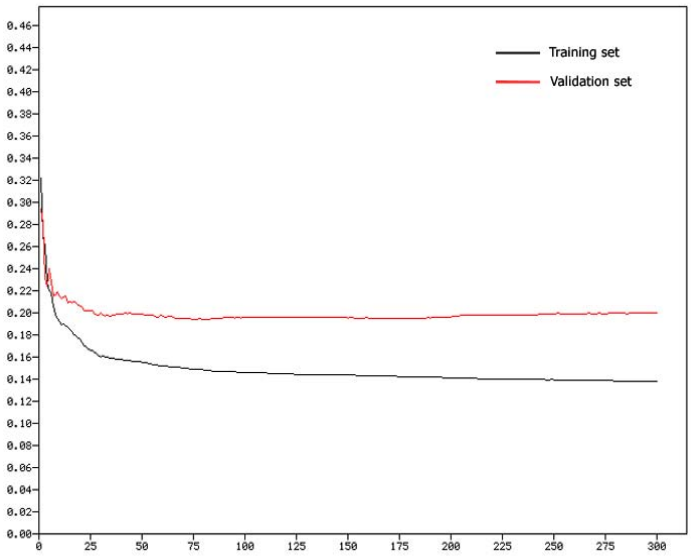
These indicators put the network on the right track in two ways. First, while investment can be equal to zero in one year and in one aggregation (either machinery or software), the more zeros we find, the more suspicious the record looks: it is very unlikely a firm will refrain totally from any investment for two consecutive years. Second, firms that invest for two years in a row in a given item might show definite patterns, different from those of firms that only choose to invest during

one of the two years. For example, suppose that the former do not invest less than the latter, on average: this will be reflected in the variance of the distribution of investment-to-sales and investment-per-employee ratios over the two years. If the learning process is helped by inserting a simple Boolean indicator that signals whether a firm reportedly invested or not during each year, it will tell true zeros from false zeros more easily: the network may deduce a requirement for consistency of this indicator with the patterns seen on related variables. For similar reasons, we also added variables that contain the sum of investments in all fields over the two years.

Figure 6 shows the MSE evaluated on the training and on the validation set and Tables 3a-3d present the results.

Figure 6

Training and validation MSE: investment



Tables 3a-3d

INVESTMENT: CONFUSION MATRIX FOR THE TRAINING SET
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	631 42.52 85.04 76.76	644 43.4 86.79 64.02	111 7.48 14.96 16.77	98 6.6 13.21 20.5	742 50
	Wrong	191 12.87 25.74 23.24	362 24.39 48.79 35.98	551 37.13 74.26 83.23	380 25.61 51.21 79.5	742 50
	Total	822 55.39	1006 67.79	662 44.61	478 32.21	1484 100.00

Correct predictions: ANN 79.65 per cent; logistic regression 69.01 per cent

INVESTMENT: CONFUSION MATRIX FOR THE VALIDATION SET
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	615 41.61 74.73 75	643 43.5 78.13 66.84	208 14.07 25.27 31.61	180 12.18 21.87 34.88	823 55.68
	Wrong	205 13.87 31.3 25	319 21.58 48.7 33.16	450 30.45 68.7 68.39	336 22.73 51.3 65.12	655 44.32
	Total	820 55.48	962 65.09	658 44.52	516 34.91	1478 100.00

Correct predictions: ANN 72.06 per cent; logistic regression 66.23 per cent

INVESTMENT: CONFUSION MATRIX FOR THE TEST SET I
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	874 58.89 75.93 84.85	954 64.29 82.88 79.17	277 18.67 24.07 61.01	197 13.27 17.12 70.61	1151 77.56
	Wrong	156 10.51 46.85 15.15	251 16.91 75.38 20.83	177 11.93 53.15 38.99	82 5.53 24.62 29.39	333 22.44
	Total	1030 69.41	1205 81.2	454 30.59	279 18.8	1484 100.00

Correct predictions: ANN 70.82 per cent; logistic regression 69.82 per cent

INVESTMENT: CONFUSION MATRIX FOR THE TEST SET II
(frequency, per cent, row per cent, column per cent)

		PREDICTED				Total
		Correct		Wrong		
		ANN	Logistic	ANN	Logistic	
TRUE	**Correct**	854 57.78 74.85 86	911 61.64 79.84 78.40	287 19.42 25.15 59.18	230 15.56 20.16 72.78	1141 77.2
	Wrong	139 9.4 41.25 14	251 16.98 74.48 21.60	198 13.4 58.75 40.82	86 5.82 25.52 27.22	337 22.8
	Total	993 67.19	1162 78.62	485 32.81	316 21.38	1478 100.00

Correct predictions: ANN 71.18 per cent; logistic regression 67.43 per cent

On the training set, 79.65 per cent of all records are correctly classified; the hit rate is 72.06 per cent on the validation set, and respectively 70.82 and 71.18 per cent on the two test sets. The correct predictions, conditional on no mistakes, range between 75 and 85 per cent depending on the dataset, both for the network and the competing logistic model. On the other hand, the network again outperforms the logistic model on the very task of error-spotting (70 per cent of wrong records, against 50 per cent for the logistic model). On the test sets, the network shows a relatively poorer performance, as anticipated;[10] it only catches, respectively, 53.15 and 58.75 per cent of the errors, still better than the logistic predictor (25 per cent in both cases).

Predictions formulated on investment are on the whole less certain than those referring to other phenomena. The average distance between target and forecast is considerably larger than in the case of employment or sales, as Figure 4 clearly shows, but the hit rate is not dramatically smaller: since all predictions smaller than 0.5 are mapped to zero and the rest to one, the hit rate is only partly affected by the fact that investment predictions are closer to the centre of the (0,1) interval than the other predictions. This can be seen clearly in the kernel approximations shown in Section 5.6 below.

5.6 An interpretation of the results in terms of density estimation

All the results are produced according to the generally accepted convention of translating numbers below 0.5 as a prediction of correctness (value 0), so that records with a network output such as 0.45 are going to be evaluated as correct: if we estimate a kernel density function of the predictions respectively of absence and presence of errors (Figures 7 and 8) to account for the shape of the empirical distribution function, we get a clearer understanding of the functioning of the networks. The excellent performance in error identification of sales is shown by the density peaks respectively located very close to zero and to one. Slightly less good, but equally satisfying, is the capability of the network for employment, which presents a negligible uncertainty in correctly identifying wrong values (as shown by the hump on the left of Figure 8). The estimated distributions confirm that investment predictions are more concentrated towards the centre of the interval (0,1): for investment, the network is apparently more cautious, whereas for employment and sales is more clear-cut in assessing the probability of a record containing mistakes.

[10] As explained in Section 5.2, in the test phase we re-introduced perturbations such as swap of neighbouring digits, or substitution of a randomly chosen digit with a randomly selected one: we expected this to lower the hit rate, acting as a sort of “stress test”.

Figure 7

Distribution of network predictions, conditional on absence of errors (0s)
(kernel density estimates, rescaled)

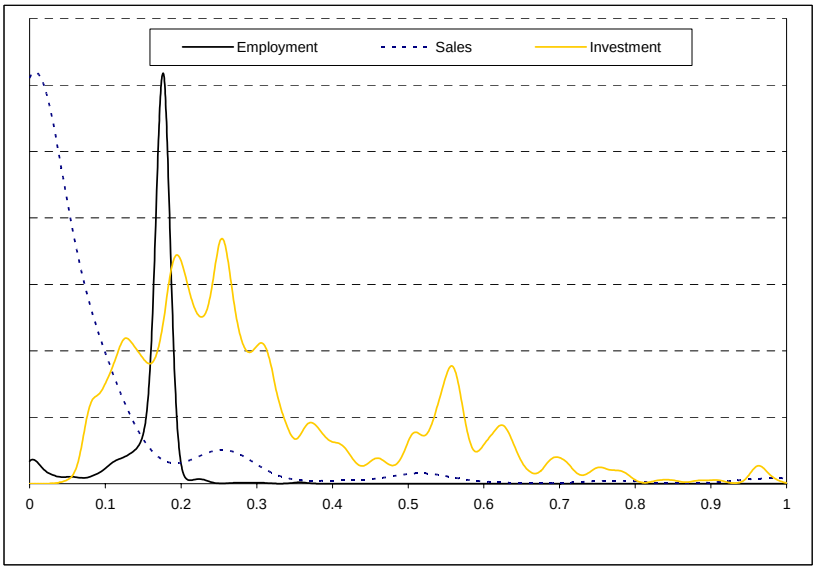
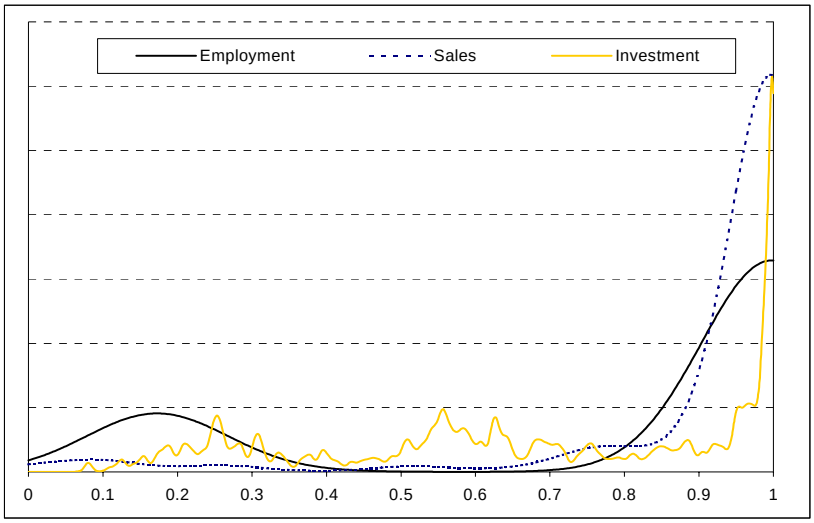


Figure 8

Distribution of network predictions, conditional on presence of errors (1s)
(kernel density estimates, rescaled)



6. Conclusions and further developments

An application of neural network models to predictive classification for data quality control is presented. Three feed-forward networks (for employment, sales and investment) are trained on a set of records from the most recent wave of the Bank of Italy's business surveys, in order to identify the presence of measurement error. The output variable is a binary proxy of unobservable

measurement error; the input variables include raw survey variables and many transforms. The networks are trained on simulated datasets, based on perturbations of the original survey dataset intended to reproduce real types of measurement error.

Common mistakes can be recognized by a properly trained neural network with a satisfying level of accuracy. The per centage of correct predictions on training datasets is 89.86 per cent for employment, 95.75 per cent for sales and 79.65 per cent for investment. Validation sets yield similar results. Out-of-sample generalizations are also good, although the test sets includes types of errors not used in the learning phase, which just slightly lower the hit rates conditional on error presence. The networks consistently outperform the logistic predictor, which is the most common alternative classification method.

The performance shown by the networks turns out to depend crucially on how prior knowledge is plugged in the architecture, and on the choice of a learning algorithm. As far as the first issue is concerned, transforms are introduced to reduce the dimensionality of the learning task: consistency problems are therefore turned into simpler range problems. Other improvements derive from simple discrete indicators of data structure that act as simple blueprints broadly followed by the network to build acceptance and rejection regions. As for the learning algorithm, the resilient propagation (Rprop) delivers largely superior results to those of the standard back-propagation (see Appendix). This probably depends on the adaptive nature of the former, which fine-tunes the learning parameter based on the data structure: a particularly useful feature in complex datasets.

In general, errors in investments appear harder to spot than those in employment and sales: a larger proportion of network predictions is actually incorrect and, when correct, less certain. This can be explained on the basis of the limited stability of investment: records with correct and wrong values can look too similar for the network to discriminate.

The effectiveness of the networks in flagging erroneous groups of variables suggests a twofold direction for further research. At the macro level, we could try to streamline the existing data editing process. As seen in Section 2, traditional selective editing procedures suffer from a low “hit rate”, which hovers around a fifth of the flagged records.

The data editing methods currently used split the records in two groups: the thoroughly edited records, the so-called “critical stream” (the most influential units) and the loosely edited records, the “non-critical stream” (the less influential units). On the “critical stream”, the integration of network flagging with traditional (score-based) flagging can be experimented.

On the “non-critical stream”, the use of ANN could become an effective editing tool in data quality assessment for micro level analysis. Neural network error predictions are most valuable here: traditional microdata quality assessment is a very complex task, largely based on expert advice, whereas the ANN learning process guarantees that units are compared to each other by

exploiting known regularities found in units flagged as correct. The flexible learning mechanism enables any useful new features of error-free units to be inserted easily into the model.

Appendix: Methodological issues

Basics on neural networks

Artificial neural networks (ANNs) have aroused growing interest in the last dozen years even outside of their native field of application, which was traditionally discriminant analysis and pattern recognition. The name recalls the very activity of the human brain, structured in collection of information, learning by training and subsequent validation; it has a suggestive flavour and may possibly raise misplaced enthusiasm for a promising new tool.

Research into ANNs dates back to the studies of McCulloch and Pitts (1943) but gained momentum only in the 1960s, as increasingly available computing power made the widespread implementation of complex algorithms possible. The seminal idea of a linear combination of observed values to separate (classify) units is found in the work of Rosenblatt (1962). The re-emergence of ANNs in the 1980s is marked by Rumelhart and McClelland's (1986) textbook. Recent references are, for example, Kröse and van der Smagt (1996) and Ripley (1996).

From an econometric point of view, ANNs can definitely be viewed as powerful prediction tools with strong connections to the classical framework of multivariate regression models. Systematization of ANN theory under the category of statistical computing started with the classical paper of Cheng and Titterington (1994); a discussion that carefully parallels ANNs and non-linear regression tools, comparing also the respective jargon, can be found in Martin and Tan (1997).

Neural networks can be fully represented by means of partially connected graphs. Units, or nodes, are called "neurons" in ANN parlance; nodes and arches are organized in layers. The first layer contains inputs, i.e. right-hand side variables; the last final layer contains outputs, i.e. left-hand side variables. Layers between the initial and the final ones are called "hidden" layers, as their values are fed from within the network and do not represent observed variables.

In a feed-forward network, information flows from input to output only, with no arches reverting to a node belonging to a preceding layer. This is perfectly intuitive as we keep constantly in mind the comparison with regressions models, to which feed-forward ANNs can be reconducted. Impulses (=values) coming from input and intermediate layers are magnified or inhibited through appropriate weights and then fed (=weighted sum) to the subsequent layer. Prior to that, the signals are processed through a function, called a "squasher" or activation, that calibrates the sensitivity of the nodes to the stimuli. Activation functions are typically sigmoidal (e.g. logistic or cumulative normal) or threshold.

"Learning" means that the output of the network is compared with the desired output by means of a loss function. If the value of the loss function is not minimum, an appropriate rule (typically the gradient method) adjusts the weights, whose values are propagated back through the network. This

way of iterating is not unique but typical, and is called back-propagation. The learning process ends when the distance between desired and predicted output is minimum, i.e. global loss is minimum.

If the desired value of the output is compared with the network for learning (=fitting) we speak of "supervised" learning. So far this is the only type of ANN that has proved of use for concrete applications, as unsupervised learning poses expectedly formidable methodological problems.
When it comes to predicting a zero-one classifying variable, as in our case, the logit modelling framework appears a natural competitor. In fact, there is perfect coincidence between a particular class of artificial neural networks, i.e. two-layer (no hidden layer), feed-forward networks with a single logistic activation function (at the output node), and multinomial logit (Ripley 1996); the enhancement offered by neural networks equipped with hidden layers lies in the fact that they can natively account for non-linearities in the hypothesized relationship. A clear comparison between feed-forward, logistic-activated ANNs and the multinomial logit model is also found in Bentz and Merunka (2000).

Basics on the applied architecture

A neural network is fully defined once we have chosen the following features: a) the network topology; b) the input transfer function; c) the weight adjustment rule; d) the output transformation function. What follows is common for the three networks we trained.

a) We chose the classical feed-forward network topology. A feed-forward network has vertices that can be numbered so that all connections go from each vertex to another with a higher number. Nodes are organized in layers with connections available only from lower-number to higher-number layers. In practice, signals flow in one direction only (from input to output). This recalls the familiar feature of ordinary regression models, where we feed values in right-hand variables and get the resulting output from the left-hand side.

Given the universal approximation property shown, for example, in Hornik, Stinchcombe and White (1989), one hidden layer is enough for our purposes. Such a network can be represented by the expression (Ripley, 1996)

$$y_l = f_l\left(\alpha_k + \sum_{j=1}^{k} w_{jk} f_j\left(\alpha_j + \sum_{i=1}^{j} w_{ij} f_i(x_i)\right)\right)$$

a possibly non-linear regression function. Here j is the number of input nodes, k the number of hidden nodes, l the number of output nodes ($l=1$ in our case). The $f(\cdot)$s are called the activation functions; the ws are the weights assigned to each node and account for the approximation of the y_l

through the network. As seen in Section 2, a network with linear input and a single logistic output can be seen as an extension of the logistic regression; they coincide if we get rid of the hidden layer.

b) The hidden layer activation function has been chosen as logistic, as this type of activation has proved able to learn and converge on our data. Also the output node has a logistic activation. A dummy value is finally obtained by taking, as usual, the raw value at the output node of the network – a number in the open interval (0,1) – and clipping it to zero if below 0.5, to 1 in the opposite case. The final output is thus represented by a dummy variable, assuming values 0 or 1 according respectively to the absence or presence of errors in any of the variables of the group.

c) the number of units in the hidden layer has been optimized by the simple procedure known as *magnitude-based pruning*: the arc whose weight is smallest (in absolute size) is removed after each training. The procedure stops as the maximum acceptable distance between MSEs before and after pruning is reached.

d) We chose resilient back-propagation (Rprop) (Riedmiller and Braun, 1993) as the rule for weight updating. This represents a major improvement on the traditional back-propagation[11] algorithm: here the updates are no longer proportional to the partial derivative, as they use an independent step size for every connection. The direction is defined only by the sign of the partial derivative. The sign-based scheme allows for the reduction of potentially distorsive influences of the derivatives' magnitude on the weight updates, and has been shown to be suitable for applications where the error is noisy. The formulae for Rprop are intuitive but a bit bulky: the learning rule for the update-values Δ^t_{ij} is

$$\Delta^t_{ij} = \begin{cases} \eta^+ \Delta^{t-1}_{ij}, \dfrac{\partial d^{t-1}}{\partial w_{ij}} \cdot \dfrac{\partial d^t}{\partial w_{ij}} > 0 \\ \eta^- \Delta^{t-1}_{ij}, \dfrac{\partial d^{t-1}}{\partial w_{ij}} \cdot \dfrac{\partial d^t}{\partial w_{ij}} < 0 \\ \Delta^{t-1}_{ij} otherwise \end{cases}$$

[11] The basics of vanilla back-propagation can be expounded briefly as follows. If we have "examples" (**x**, **t**) (**t** fitting **y**) and the output of the network is $\boldsymbol{y}=f(\mathbf{x};\mathbf{w})$ the parameter vector **w** should minimize $d(\mathbf{t},\mathbf{y})$, where d is an average distance function. If this distance is convex, e.g. a sum of squares, the "steepest descent" method can be used for updating the weights: the rule is of the type

$$w^{(n+1)} = w^{(n)} - \eta \frac{\partial d}{\partial w}\Big|_{w=w^{(n)}} + \zeta (w^{(n)} - w^{(n-1)})$$

Where η marks the rate of adjustment $\eta \leq 1$ and is called the "learning rate". ζ, called "momentum", is an optional term which allows additional flexibility in updating during the iterations and helps to go past local minima. This procedure is in fact a least square fit achieved through an iterative algorithm (due to the non-linear structure of the model).

Back-propagation, being basically a gradient convergence method bound by the choice of a fixed "learning rate" and making use of the first derivative only, can be rather slow, as shown in the literature. Moreover, it is rather prone to get stuck in paralysis (if changes in the weights become too small) or to get trapped into local minima. A remedy to these drawbacks seems to have been found in Rprop.

and, consequently, the learning rule for the weights is described by $w_{ij}^{t+1}=w_{ij}^{t}+\Delta w_{ij}^{t}$, where

$$\Delta w_{ij}^{t} = \begin{cases} -\Delta_{ij}^{t}, & \dfrac{\partial d^{t}}{\partial w_{ij}} > 0 \\ +\Delta_{ij}^{t}, & \dfrac{\partial d^{t}}{\partial w_{ij}} < 0 \\ 0 & otherwise \end{cases}$$

Whenever a change of sign occurs for the partial derivative of a weight, (signalling that the last update was too big and the procedure got trapped in a local minimum) the update-value is decreased by a fixed factor. If the derivative retains its sign instead, the update-value is increased to speed convergence. Once the update value has been adapted, the weights are updated in turn by the intuitive rule: if the derivative is positive, the weight is decreased, and vice versa. The Rprop approach has triggered a slew of research to improve this algorithm further: new techniques like *QRprop*, *IRprop*, *GRprop* have been proposed (see, for example, Aristoklis et al. 2005), although they are not yet implemented in the software package we used.

The software

The experiments were conducted by means of JavaNNS software package (Java Neural Networks Simulator, version 1.1) in a PC Windows NT (ver. 4.0) architecture. For most computationally intensive tasks, the UNIX version of SNNS (Stuttgart Neural Network Simulator) package was run on a RISC system.

JavaNNS is a simulator for neural networks developed at the Wilhelm-Schickard-Institute for Computer Science (WSI) in Tübingen, Germany. It is based on the SNNS 4.2 kernel, with a new graphical user interface written in Java. Currently, JavaNNS is distributed by the University of Tübingen only as a binary file. It is not public domain, but is available free of charge. All relevant information on this package can be found at the Web address:

http://www-ra.informatik.uni-tuebingen.de/software/JavaNNS/welcome_e.html

References

Aristoklis, D. A., Magoulasa, G. D. and Vrahatisb, M. N. (2005), New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing*, 64, 253–270.

Barcaroli, G. and D'Aurizio, L. (1997), Evaluating editing procedures: the simulation approach, *Working Paper, Conference of European Statisticians, Work Session on Statistical Data Editing,* Prague.

Battipaglia, P. (2002), Selective editing to increase efficiency in survey data processing – An application to the Bank of Italy's Business Survey on Industrial Firms, *Irving Fisher Committee Bulletin*, n. 13, December.

Bentz, Y. and Merunka, D. (2000), Neural networks and the multinomial logit for brand choice modelling: a hybrid approach, *Journal of Forecasting*, 19, 177-200.

Biancotti, C. and Tartaglia-Polcini, R. (2005), Artificial Neural Networks for Data Editing, *Irving Fisher Committee Bulletin,* n. 21, July, 99-107.

Breiman, L. (1994), Comment of Neural networks: a review from a statistical perspective, by Cheng, B. And Titterington, M., *Statistical Science*, vol. 9 n. 1, 2-54.

Breiman, L. (2001), Statistical modeling: the two cultures, *Statistical Science*, vol. 16 n. 3, 199-231.

Brackstone, G. (1999), Managing data quality in a statistical agency, *Survey Methodology*, vol. 25 n. 2, 139-149.

Cheng, B. and Titterington, D. M. (1994), Neural networks: a review from a statistical perspective (with discussion), *Statistical Science*, 9, 2-54.

Chinnappa, N. et al. (1990), "Macro editing at Statistics Canada", unpublished report of the Statistics Canada working group on strategies for macro editing, prepared for the Statistics Canada Advisory Committee on statistical methods (January), Ottawa: Statistics Canada.

Graepel, T. et al. (2001), Neural networks in economics: background, applications and new developments. Department of computer science, Technical University of Berlin.

Granquist, L. and Kovar, J.G. (1997), "Editing of survey data: how much is enough?" in: Survey measurement and process quality, edited by Lyberg, Biemer et al. New York: Wiley.

Hedlin, D. (2003), Score functions to reduce Business Survey Editing at the U.K. Office for National Statistics, Journal of Official Statistics, Vol. 19 n. 2, pp. 177-199.

Hornik, K., Stinchcombe, M. and White, H. (1989), Multi-layer feedforward networks as universal approximators, *Neural Networks*, 2, 359-366.

Kröse, B. and van der Smagt, P. (1996), An introduction to neural networks. Eight edition. Amsterdam: University of Amsterdam.

Larsen, B. and Madsen, B. (1999), Error identification and imputations with neural networks, UN/ECE Work Session on Statistical Data Editing, Working Paper 26.

Martin, V. L. and Tan, C. (1997), Artificial neural networks, in: Creedy, J. and V. L. Martin (eds.), *Nonlinear economic models*, Cheltenham: Edward Elgar.

Rohwer, R, Wynne-Jones, M. and Wysotzki, F. (1995), Neural Networks. In Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (editors), Machine learning, neural and statistical classification. Hertfordshire: Ellis Horwood.

Nordbotten, S. (1995), Editing statistical records by neural networks, *Journal of Official Statistics* Vol. 11 n. 4, 391-411.

Nordbotten, S. (1996), Editing and imputation by means of neural networks, *Statistical Journal of the United Nations Economic Commission for Europe*, 119-129.

Riedmiller, M. and Braun, H. (1993), "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm", Proceedings of the IEEE International Conference on Neural Networks 1993, San Francisco: IEEE.

Ripley, B.D. (1996), Pattern recognition and neural networks. Cambridge: Cambridge University Press.

Rivière, P. (2002), General data editing tools are often unsuitable to use in complex business surveys: why?, Conference of European Statisticians, UNECE Work Session on Data Editing, Helsinki, Working Paper n. 30.

Roddick L. H. (1995), "Data Editing using Neural Networks, *UN/ECE Work Session on Statistical Data Editing*, Athens, Greece.

Rosenblatt, F. (1962), Principles of Neurodynamics. Washington: Spartan books.

Rumelhart, D.E. and MacClelland, J.L. (1986), Parallel distributed processing, Cambridge: MIT Press.

RECENTLY PUBLISHED "TEMI" (*).

N. 589 – *An empirical analysis of national differences in the retail bank interest rates of the euro area*, by M. Affinito and F. Farabullini (May 2006).

N. 590 – *Imperfect knowledge, adaptive learning and the bias against activist monetary policies*, by Alberto Locarno (May 2006).

N. 591 – *The legacy of history for economic development: the case of Putnam's social capital,* by G. de Blasio and G. Nuzzo (May 2006).

N. 592 – *L'internazionalizzazione produttiva italiana e i distretti industriali: un'analisi degli investimenti diretti all'estero,* by Stefano Federico (May 2006).

N. 593 – *Do market-based indicators anticipate rating agencies? Evidence for international banks,* by Antonio Di Cesare (May 2006).

N. 594 – *Entry regulations and labor market outcomes: Evidence from the Italian retail trade sector,* by Eliana Viviano (May 2006).

N. 595 – *Revisiting the empirical evidence on firms' money demand,* by Francesca Lotti and Juri Marcucci (May 2006).

N. 596 – *Social interactions in high school: Lesson from an earthquake*, by Piero Cipollone and Alfonso Rosolia (September 2006).

N. 597 – *Determinants of long-run regional productivity: The role of R&D, human capital and public infrastructure*, by Raffaello Bronzini and Paolo Piselli (September 2006).

N. 598 – *Overoptimism and lender liability in the consumer credit market*, by Elisabetta Iossa and Giuliana Palumbo (September 2006).

N. 599 – *Bank's riskiness over the business cycle: A panel analysis on Italian intermediaries*, by Mario Quagliariello (september 2006)

N. 600 – *People I know: Workplace networks and job search outcomes*, by Federico Cingano and Alfonso Rosolia (September 2006).

N. 601 – *Bank profitability and the business cycle*, by Ugo Albertazzi and Leonardo Gambacorta (September 2006).

N. 602 – *Scenario based principal component value-at-risk: An application to Italian banks' interest rate risk exposure*, by Roberta Fiori and Simonetta Iannotti (September 2006).

N. 603 – *A dual-regime utility model for poverty analysis*, by Claudia Biancotti (September 2006).

N. 604 – *The political economy of investor protection*, by Pietro Tommasino (December 2006).

N. 605 – *Search in thick markets: Evidence from Italy*, by Sabrina Di Addario (December 2006).

N. 606 – *The transmission of monetary policy shocks from the US to the euro area*, by S. Neri and A. Nobili (December 2006).

N. 607 – *What does a technology shock do? A VAR analysis with model-based sign restrictions*, by L. Dedola and S. Neri (December 2006).

N. 608 – *Merge and compete: Strategic incentives for vertical integration*, by Filippo Vergara Caffarelli (December 2006).

N. 609 – *Real-time determinants of fiscal policies in the euro area: Fiscal rules, cyclical conditions and elections*, by Roberto Golinelli and Sandro Momigliano (December 2006).

N. 610 – *L'under-reporting della ricchezza finanziaria nell'indagine sui bilanci delle famiglie*, by Leandro D'Aurizio, Ivan Faiella, Stefano Iezzi, Andrea Neri (December 2006).

N. 611 – *La polarizzazione territoriale del prodotto pro capite: un'analisi del caso italiano sulla base di dati provinciali* by Stefano Iezzi (December 2006).

(*) Requests for copies should be sent to:
Banca d'Italia – Servizio Studi – Divisione Biblioteca e pubblicazioni – Via Nazionale, 91 – 00184 Rome (fax 0039 06 47922059). They are available on the Internet www.bancaditalia.it.

2000

P. ANGELINI, *Are banks risk-averse? Intraday timing of the operations in the interbank market*, Journal of Money, Credit and Banking, Vol. 32 (1), pp. 54-73, **TD No. 266 (April 1996).**

F. DRUDI and R. GIORDANO, *Default Risk and optimal debt management,* Journal of Banking and Finance, Vol. 24 (6), pp. 861-891, **TD No. 278 (September 1996)**.

F. DRUDI and R. GIORDANO, *Wage indexation, employment and inflation,* Scandinavian Journal of Economics, Vol. 102 (4), pp. 645-668, **TD No. 292 (December 1996)**.

F. DRUDI and A. PRATI, *Signaling fiscal regime sustainability,* European Economic Review, Vol. 44 (10), pp. 1897-1930, **TD No. 335 (September 1998)**.

F. FORNARI and R. VIOLI, *The probability density function of interest rates implied in the price of options,* in: R. Violi, (ed.) , Mercati dei derivati, controllo monetario e stabilità finanziaria, Il Mulino, Bologna, **TD No. 339 (October 1998)**.

D. J. MARCHETTI and G. PARIGI, *Energy consumption, survey data and the prediction of industrial production in Italy,* Journal of Forecasting, Vol. 19 (5), pp. 419-440, **TD No. 342 (December 1998)**.

A. BAFFIGI, M. PAGNINI and F. QUINTILIANI, *Localismo bancario e distretti industriali: assetto dei mercati del credito e finanziamento degli investimenti,* in: L.F. Signorini (ed.), Lo sviluppo locale: un'indagine della Banca d'Italia sui distretti industriali, pp. 237-256, Meridiana Libri, **TD No. 347 (March 1999)**.

F. LIPPI, *Median voter preferences, central bank independence and conservatism*, Public Choice, v. 105, 3-4, pp. 323-338 **TD No. 351 (April 1999)**.

A. SCALIA and V. VACCA, *Does market transparency matter? A case study*, in: Market Liquidity: Research Findings and Selected Policy Implications*,* Basel, Bank for International Settlements, **TD No. 359 (October 1999)**.

F. SCHIVARDI, *Rigidità nel mercato del lavoro, disoccupazione e crescita*, Giornale degli economisti e Annali di economia, Vol. 59 (1), pp. 115-141, **TD No. 364 (December 1999)**.

G. BODO, R. GOLINELLI and G. PARIGI, *Forecasting industrial production in the euro area,* Empirical Economics, Vol. 25 (4), pp. 541-561, **TD No. 370 (March 2000)**.

F. ALTISSIMO, D. J. MARCHETTI and G. P. ONETO, *The Italian business cycle: Coincident and leading indicators and some stylized facts,* Giornale degli economisti e Annali di economia, Vol. 60 (2), pp. 147-220, **TD No. 377 (October 2000)**.

C. MICHELACCI and P. ZAFFARONI, *(Fractional) Beta convergence*, Journal of Monetary Economics, Vol. 45 (1), pp. 129-153, **TD No. 383 (October 2000)**.

R. DE BONIS and A. FERRANDO*, The Italian banking structure in the nineties: Testing the multimarket contact hypothesis*, Economic Notes, Vol. 29 (2), pp. 215-241, **TD No. 387 (October 2000)**.

S. SIVIERO and D. TERLIZZESE*, La previsione macroeconomica: alcuni luoghi comuni da sfatare*, Rivista italiana degli economisti, v. 5, 2, pp. 291-322, **TD No. 395 (February 2001)**.

G. DE BLASIO and F. MINI*, Seasonality and capacity: An application to Italy*, IMF Working Paper, 80, **TD No. 403 (June 2001)**.

2001

M. CARUSO, *Stock prices and money velocity: A multi-country analysis,* Empirical Economics, Vol. 26 (4), pp. 651-672, **TD No. 264 (February 1996)**.

P. CIPOLLONE and D. J. MARCHETTI, *Bottlenecks and limits to growth: A multisectoral analysis of Italian industry*, Journal of Policy Modeling, Vol. 23 (6), pp. 601-620, **TD No. 314 (August 1997)**.

P. CASELLI, *Fiscal consolidations under fixed exchange rates,* European Economic Review, Vol. 45 (3), pp. 425-450, **TD No. 336 (October 1998)**.

F. ALTISSIMO and G. L. VIOLANTE, *The non-linear dynamics of output and unemployment in the US,* Journal of Applied Econometrics, Vol. 16 (4), pp. 461-486, **TD No. 338 (October 1998)**.

F. NUCCI and A. F. POZZOLO, *Investment and the exchange rate: An analysis with firm-level panel data*, European Economic Review, Vol. 45 (2), pp. 259-283, **TD No. 344 (December 1998)**.

A. ZAGHINI, *Fiscal adjustments and economic performing: A comparative study*, Applied Economics, Vol. 33 (5), pp. 613-624, **TD No. 355 (June 1999)**.

L. GAMBACORTA, *On the institutional design of the European monetary union: Conservatism, stability pact and economic shocks,* Economic Notes, Vol. 30 (1), pp. 109-143, **TD No. 356 (June 1999)**.

P. FINALDI RUSSO and P. ROSSI, *Credit costraints in italian industrial districts,* Applied Economics, Vol. 33 (11), pp. 1469-1477, **TD No. 360 (December 1999)**.

A. CUKIERMAN and F. LIPPI, *Labor markets and monetary union: A strategic analysis,* Economic Journal, Vol. 111 (473), pp. 541-565, **TD No. 365 (February 2000)**.

G. PARIGI and S. SIVIERO, *An investment-function-based measure of capacity utilisation, potential output and utilised capacity in the Bank of Italy's quarterly model*, Economic Modelling, Vol. 18 (4), pp. 525-550, **TD No. 367 (February 2000)**.

P. CASELLI, P. PAGANO and F. SCHIVARDI, *Investment and growth in Europe and in the United States in the nineties,* Rivista di politica economica, v. 91, 10, pp. 3-35, **TD No. 372 (March 2000)**.

F. BALASSONE and D. MONACELLI, *Emu fiscal rules: Is there a gap?,* in: M. Bordignon and D. Da Empoli (eds.), Politica fiscale, flessibilità dei mercati e crescita*,* Milano*,* Franco Angeli, **TD No. 375 (July 2000)**.

A. B. ATKINSON and A. BRANDOLINI, *Promise and pitfalls in the use of "secondary" data-sets: Income inequality in OECD countries as a case study*, Journal of Economic Literature, Vol. 39 (3), pp. 771-799, **TD No. 379 (October 2000)**.

D. FOCARELLI and A. F. POZZOLO, *The patterns of cross-border bank mergers and shareholdings in OECD countries*, Journal of Banking and Finance, Vol. 25 (12), pp. 2305-2337, **TD No. 381 (October 2000)**.

M. SBRACIA and A. ZAGHINI, *Expectations and information in second generation currency crises models*, Economic Modelling, Vol. 18 (2), pp. 203-222, **TD No. 391 (December 2000)**.

F. FORNARI and A. MELE, *Recovering the probability density function of asset prices using GARCH as diffusion approximations,* Journal of Empirical Finance, Vol. 8 (1), pp. 83-110**, TD No. 396 (February 2001)**.

P. CIPOLLONE, *La convergenza dei salari dell'industria manifatturiera in Europa,* Politica economica, Vol. 17 (1), pp. 97-125, **TD No. 398 (February 2001)**.

E. BONACCORSI DI PATTI and G. GOBBI, *The changing structure of local credit markets: Are small businesses special?*, Journal of Banking and Finance, Vol. 25 (12), pp. 2209-2237, **TD No. 404 (June 2001)**.

L. DEDOLA and S. LEDUC, *Why is the business-cycle behaviour of fundamentals alike across exchange-rate regimes?,* International Journal of Finance and Economics, v. 6, 4, pp. 401-419, **TD No. 411 (August 2001)**.

M. PAIELLA, *Limited Financial Market Participation: a Transaction Cost-Based Explanation,* IFS Working Paper, 01/06, **TD No. 415 (August 2001)**.

G. MESSINA, *Per un federalismo equo e solidale: obiettivi e vincoli per la perequazione regionale in Italia*,, Studi economici, Vol. 56 (73), pp. 131-148, **TD No. 416 (August 2001)**.

L GAMBACORTA *Bank-specific characteristics and monetary policy transmission: the case of Italy,* ECB Working Paper, 103, **TD No. 430 (December 2001)**.

F. ALTISSIMO, A. BASSANETTI, R. CRISTADORO, M. FORNI, M. LIPPI, L. REICHLIN and G. VERONESE *A real time coincident indicator of the euro area business cycle,* CEPR Discussion Paper, 3108, **TD No. 436 (December 2001)**.

A. GERALI and F. LIPPI, *On the "conquest" of inflation,* CEPR Discussion Paper, 3101, **TD No. 444 (July 2002)**.

L. GUISO and M. PAIELLA, *Risk aversion, wealth and background risk,* CEPR Discussion Paper, 2728, **TD No. 483 (September 2003)**.

2002

R. CESARI and F. PANETTA, *The performance of italian equity fund*, Journal of Banking and Finance, Vol. 26 (1), pp. 99-126, **TD No. 325 (January 1998)**.

F. ALTISSIMO, S. SIVIERO and D. TERLIZZESE, *How deep are the deep parameters?,* Annales d'Economie et de Statistique, (67/68), pp. 207-226, **TD No. 354 (June 1999)**.

F. FORNARI, C. MONTICELLI, M. PERICOLI and M. TIVEGNA, *The impact of news on the exchange rate of the lira and long-term interest rates,* Economic Modelling, Vol. 19 (4), pp. 611-639, **TD No. 358 (October 1999**).

D. FOCARELLI, F. PANETTA and C. SALLEO, *Why do banks merge?,* Journal of Money, Credit and Banking, Vol. 34 (4), pp. 1047-1066, **TD No. 361 (December 1999)**.

D. J. MARCHETTI, *Markup and the business cycle: Evidence from Italian manufacturing branches,* Open Economies Review, Vol. 13 (1), pp. 87-103, **TD No. 362 (December 1999**).

F. BUSETTI, *Testing for (common) stochastic trends in the presence of structural break*, Journal of Forecasting, Vol. 21 (2), pp. 81-105, **TD No. 385 (October 2000)**.

F. LIPPI, *Revisiting the Case for a Populist Central Banker,* European Economic Review*,* Vol. 46 (3), pp. 601-612, **TD No. 386 (October 2000)**.

F. PANETTA, *The stability of the relation between the stock market and macroeconomic forces*, Economic Notes, Vol. 31 (3), pp. 417-450, **TD No. 393 (February 2001)**.

G. GRANDE and L. VENTURA, *Labor income and risky assets under market incompleteness: Evidence from Italian data,* Journal of Banking and Finance, Vol. 26 (2-3), pp. 597-620, **TD No. 399 (March 2001)**.

A. BRANDOLINI, P. CIPOLLONE and P. SESTITO, *Earnings dispersion, low pay and household poverty in Italy, 1977-1998,* in D. Cohen, T. Piketty and G. Saint-Paul (eds.), The Economics of Rising Inequalities, Oxford, Oxford University Press, **TD No. 427 (November 2001)**.

E. GAIOTTI and A. GENERALE, *Does monetary policy have asymmetric effects? A look at the investment decisions of Italian firms,* Giornale degli economisti e annali di economia, v. 61, 1, pp. 29-60, **TD No. 429 (December 2001)**.

G. M. TOMAT, *Durable goods, price indexes and quality change: An application to automobile prices in Italy, 1988-1998,* ECB Working Paper, 118, **TD No. 439 (March 2002)**.

A. PRATI and M. SBRACIA, *Currency crises and uncertainty about fundamentals,* IMF Working Paper, 3, **TD No. 446 (July 2002)**.

L. CANNARI and G. D'ALESSIO, *La distribuzione del reddito e della ricchezza nelle regioni italiane*, Rivista Economica del Mezzogiorno, Vol. 16 (4), pp. 809-847, Il Mulino, **TD No. 482 (June 2003)**.

2003

L. GAMBACORTA, *Asymmetric bank lending channels and ECB monetary policy,* Economic Modelling, Vol. 20, 1, pp. 25-46, **TD No. 340 (October 1998)**.

F. SCHIVARDI, *Reallocation and learning over the business cycle,* European Economic Review, Vol. 47 (1), pp. 95-111, **TD No. 345 (December 1998)**.

P. CASELLI, P. PAGANO and F. SCHIVARDI, *Uncertainty and slowdown of capital accumulation in Europe*, Applied Economics, Vol. 35 (1), pp. 79-89, **TD No. 372 (March 2000).**

F. LIPPI, *Strategic monetary policy with non-atomistic wage setters,* Review of Economic Studies, v. 70, 4, pp. 909-919, **TD No. 374 (June 2000)**.

P. ANGELINI and N. CETORELLI*, The effect of regulatory reform on competition in the banking industry,* Journal of Money, Credit and Banking, Vol. 35, 5, pp. 663-684, **TD No. 380 (October 2000)**.

P. PAGANO and G. FERRAGUTO, Endogenous growth with intertemporally dependent preferences, Contribution to Macroeconomics, Vol. 3 (1), pp. 1-38, **TD No. 382 (October 2000).**

P. PAGANO and F. SCHIVARDI, *Firm size distribution and growth,* Scandinavian Journal of Economics, Vol. 105 (2), pp. 255-274, **TD No. 394 (February 2001)**.

M. PERICOLI and M. SBRACIA, *A Primer on Financial Contagion,* Journal of Economic Surveys, Vol. 17 (4), pp. 571-608, **TD No. 407 (June 2001)**.

M. SBRACIA and A. ZAGHINI, *The role of the banking system in the international transmission of shocks,* World Economy, Vol. 26 (5), pp. 727-754, **TD No. 409 (June 2001)**.

L. GAMBACORTA, *The Italian banking system and monetary policy transmission: evidence from bank level data*, in: I. Angeloni, A. Kashyap and B. Mojon (eds.), Monetary Policy Transmission in the Euro Area, Cambridge University Press, **TD No. 430 (December 2001).**

M. EHRMANN, L. GAMBACORTA, J. MARTÍNEZ PAGÉS, P. SEVESTRE and A. WORMS, *Financial systems and the role of banks in monetary policy transmission in the euro area,* in: I. Angeloni, A. Kashyap and

B. Mojon (eds.), Monetary Policy Transmission in the Euro Area, Cambridge, Cambridge University Press, **TD No. 432 (December 2001)**.

F. SPADAFORA, *Official bailouts, moral hazard and the "Specialtiy" of the international interbank market*, Emerging Markets Review, Vol. 4 (2), pp. 165-196, **TD No. 438 (March 2002)**.

D. FOCARELLI and F. PANETTA, *Are mergers beneficial to consumers? Evidence from the market for bank deposits*, American Economic Review, Vol. 93 (4), pp. 1152-1172, **TD No. 448 (July 2002)**.

E.VIVIANO, *Un'analisi critica delle definizioni di disoccupazione e partecipazione in Italia,* Politica Economica, Vol. 19 (1), pp. 161-190, **TD No. 450 (July 2002)**.

M. PAGNINI, *Misura e determinanti dell'agglomerazione spaziale nei comparti industriali in Italia*, Rivista di Politica Economica, Vol. 93 (3-4), pp. 149-196, **TD No. 452 (October 2002)**.

F. PANETTA, *Evoluzione del sistema bancario e finanziamento dell'economia nel Mezzogiorno,* Moneta e credito, v. 56, 222, pp. 127-160, **TD No. 467 (March 2003)**.

F. BUSETTI and A. M. ROBERT TAYLOR, *Testing against stochastic trend and seasonality in the presence of unattended breaks and unit roots*, Journal of Econometrics, Vol. 117 (1), pp. 21-53, **TD No. 470 (March 2003)**.

P. ZAFFARONI, *Testing against stochastic trend and seasonality in the presence of unattended breaks and unit roots,* Journal of Econometrics, v. 115, 2, pp. 199-258, **TD No. 472 (June 2003)**.

E. BONACCORSI DI PATTI, G. GOBBI and P. E. MISTRULLI, *Sportelli e reti telematiche nella distribuzione dei servizi bancari,* Banca impresa società, v. 2, 2, pp. 189-209, **TD No. 508 (July 2004)**.

2004

P. ANGELINI and N. CETORELLI, *Gli effetti delle modifiche normative sulla concorrenza nel mercato creditizio,* in F. Panetta (eds.), Il sistema bancario negli anni novanta: gli effetti di una trasformazione, Bologna, il Mulino, **TD No. 380 (October 2000)**.

P. CHIADES and L. GAMBACORTA, *The Bernanke and Blinder model in an open economy: The Italian case,* German Economic Review, Vol. 5 (1), pp. 1-34, **TD No. 388 (December 2000)**.

M. BUGAMELLI and P. PAGANO, *Barriers to Investment in ICT*, Applied Economics, Vol. 36 (20), pp. 2275-2286, **TD No. 420 (October 2001)**.

F. BUSETTI, *Preliminary data and econometric forecasting: An application with the Bank of Italy quarterly model*, CEPR Discussion Paper, 4382, **TD No. 437 (December 2001)**.

A. BAFFIGI, R. GOLINELLI and G. PARIGI, *Bridge models to forecast the euro area GDP*, International Journal of Forecasting, Vol. 20 (3), pp. 447-460,**TD No. 456 (December 2002)**.

D. AMEL, C. BARNES, F. PANETTA and C. SALLEO, *Consolidation and Efficiency in the Financial Sector: A Review of the International Evidence*, Journal of Banking and Finance, Vol. 28 (10), pp. 2493-2519, **TD No. 464 (December 2002).**

M. PAIELLA, *Heterogeneity in financial market participation: Appraising its implications for the C-CAPM*, Review of Finance, Vol. 8, 3, pp. 445-480, **TD No. 473 (June 2003)**.

F. CINGANO and F. SCHIVARDI, *Identifying the sources of local productivity growth,* Journal of the European Economic Association, Vol. 2 (4), pp. 720-742, **TD No. 474 (June 2003)**.

E. BARUCCI, C. IMPENNA and R. RENÒ, *Monetary integration, markets and regulation*, Research in Banking and Finance, (4), pp. 319-360, **TD No. 475 (June 2003)**.

G. ARDIZZI, *Cost efficiency in the retail payment networks: first evidence from the Italian credit card system*, Rivista di Politica Economica, Vol. 94, (3), pp. 51-82, **TD No. 480 (June 2003)**.

E. BONACCORSI DI PATTI and G. DELL'ARICCIA, *Bank competition and firm creation*, Journal of Money Credit and Banking, Vol. 36 (2), pp. 225-251, **TD No. 481 (June 2003)**.

R. GOLINELLI and G. PARIGI, *Consumer sentiment and economic activity: a cross country comparison*, Journal of Business Cycle Measurement and Analysis, Vol. 1 (2), pp. 147-170, **TD No. 484 (September 2003)**.

L. GAMBACORTA and P. E. MISTRULLI, *Does bank capital affect lending behavior?*, Journal of Financial Intermediation, Vol. 13 (4), pp. 436-457, **TD No. 486 (September 2003)**.

F. SPADAFORA, *Il pilastro privato del sistema previdenziale: il caso del Regno Unito,* Economia Pubblica, 34, (5), pp. 75-114, **TD No. 503** (**June 2004**).

C. BENTIVOGLI and F. QUINTILIANI, *Tecnologia e dinamica dei vantaggi comparati: un confronto fra quattro regioni italiane*, in C. Conigliani (eds.), *Tra sviluppo e stagnazione: l'economia dell'Emilia-Romagna*, Bologna, Il Mulino, **TD No. 522 (October 2004)**.

G. GOBBI and F. LOTTI, *Entry decisions and adverse selection: an empirical analysis of local credit markets,* Journal of Financial services Research, Vol. 26 (3), pp. 225-244, **TD No. 535 (December 2004)**.

E. GAIOTTI and F. LIPPI, *Pricing behavior and the introduction of the euro:evidence from a panel of restaurants*, Giornale degli Economisti e Annali di Economia, 2004, Vol. 63, (3/4), pp. 491-526, **TD No. 541 (February 2005)**.

2005

L. DEDOLA and F. LIPPI, *The monetary transmission mechanism: Evidence from the industries of 5 OECD countries,* European Economic Review, 2005, Vol. 49, (6), pp. 1543-1569, **TD No. 389 (December 2000)**.

D. J. MARCHETTI and F. NUCCI, *Price stickiness and the contractionary effects of technology shocks*. European Economic Review, v. 49, pp. 1137-1164, **TD No. 392 (February 2001)**.

G. CORSETTI, M. PERICOLI and M. SBRACIA, *Some contagion, some interdependence: More pitfalls in tests of financial contagion,* Journal of International Money and Finance, v. 24, 8, pp. 1177-1199, **TD No. 408 (June 2001)**.

GUISO L., L. PISTAFERRI and F. SCHIVARDI, *Insurance within the firm*. Journal of Political Economy, 113, pp. 1054-1087, **TD No. 414** (**August 2001**)

R. CRISTADORO, M. FORNI, L. REICHLIN and G. VERONESE, *A core inflation indicator for the euro area,* Journal of Money, Credit, and Banking, v. 37, 3, pp. 539-560, **TD No. 435 (December 2001)**.

F. ALTISSIMO, E. GAIOTTI and A. LOCARNO, *Is money informative? Evidence from a large model used for policy analysis,* Economic & Financial Modelling, v. 22, 2, pp. 285-304, **TD No. 445 (July 2002)**.

G. DE BLASIO and S. DI ADDARIO, *Do workers benefit from industrial agglomeration?* Journal of regional Science, Vol. 45, (4), pp. 797-827, **TD No. 453 (October 2002).**

R. TORRINI, *Cross-country differences in self-employment rates: The role of institutions*, Labour Economics, V. 12, 5, pp. 661-683, **TD No. 459 (December 2002)**.

A. CUKIERMAN and F. LIPPI, *Endogenous monetary policy with unobserved potential output,* Journal of Economic Dynamics and Control, v. 29, 11, pp. 1951-1983, **TD No. 493 (June 2004)**.

M. OMICCIOLI, *Il credito commerciale: problemi e teorie,* in L. Cannari, S. Chiri e M. Omiccioli (eds.), *Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia*, Bologna, Il Mulino, **TD No. 494 (June 2004)**.

L. CANNARI, S. CHIRI and M. OMICCIOLI, *Condizioni di pagamento e differenziazione della clientela,* in L. Cannari, S. Chiri e M. Omiccioli (eds.), *Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia*, Bologna, Il Mulino, **TD No. 495 (June 2004)**.

P. FINALDI RUSSO and L. LEVA, *Il debito commerciale in Italia: quanto contano le motivazioni finanziarie?,* in L. Cannari, S. Chiri e M. Omiccioli (eds.), *Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia*, Bologna, Il Mulino, **TD No. 496 (June 2004)**.

A. CARMIGNANI, *Funzionamento della giustizia civile e struttura finanziaria delle imprese: il ruolo del credito commerciale,* in L. Cannari, S. Chiri e M. Omiccioli (eds.), *Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia*, Bologna, Il Mulino, **TD No. 497 (June 2004)**.

G. DE BLASIO, *Credito commerciale e politica monetaria: una verifica basata sull'investimento in scorte,* in L. Cannari, S. Chiri e M. Omiccioli (eds.), *Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia*, Bologna, Il Mulino, **TD No. 498 (June 2004)**.

G. DE BLASIO, *Does trade credit substitute bank credit? Evidence from firm-level data.* Economic notes, Vol. 34 (1), pp. 85-112, **TD No. 498 (June 2004).**

A. DI CESARE, *Estimating Expectations of Shocks Using Option Prices,* The ICFAI Journal of Derivatives Markets, Vol. 2, (1), pp. 42-53, **TD No. 506 (July 2004).**

M. BENVENUTI and M. GALLO, *Il ricorso al "factoring" da parte delle imprese italiane,* in L. Cannari, S. Chiri e M. Omiccioli (eds.), *Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia*, Bologna, Il Mulino, **TD No. 518 (October 2004)**.

L. CASOLARO and L. GAMBACORTA, *Redditività bancaria e ciclo economico,* Bancaria, v. 61, 3, pp. 19-27, **TD No. 519 (October 2004)**.

F. PANETTA, F. SCHIVARDI and M. SHUM, *Do mergers improve information? Evidence from the loan market,* CEPR Discussion Paper, 4961, **TD No. 521 (October 2004)**.

P. DEL GIOVANE and R. SABBATINI, *La divergenza tra inflazione rilevata e percepita in Italia,* Bologna, Il Mulino, **TD No. 532 (December 2004).**

R. TORRINI, *Quota dei profitti e redditività del capitale in Italia: un tentativo di interpretazione*, Politica economica, v. 21, pp. 7-42, **TD No. 551 (June 2005)**.

M. OMICCIOLI, *Il credito commerciale come "collateral"*, in L. Cannari, S. Chiri, M. Omiccioli (eds.), Imprese o intermediari? Aspetti finanziari e commerciali del credito tra imprese in Italia, Bologna, il Mulino, **TD No. 553 (June 2005)**.

L. CASOLARO, L. GAMBACORTA and L. GUISO, *Regulation, formal and informal enforcement and the development of the household loan market. Lessons from Italy,* in Bertola G., Grant C. and Disney R. (eds.) The Economics of Consumer Credit: European Experience and Lessons from the US, Boston, MIT Press, **TD No. 560 (September 2005)**.

S. DI ADDARIO and E. PATACCHINI, *Wages and the city: The italian case,* University of Oxford, Department of Economics. Discussion Paper, 243, **TD No. 570 (January 2006)**.

P. ANGELINI and F. LIPPI, *Did inflation really soar after the euro changeover? Indirect evidence from ATM withdrawals,* CEPR Discussion Paper, 4950, **TD No. 581 (March 2006)**.

2006

C. BIANCOTTI, *A polarization of inequality? The distribution of national Gini coefficients 1970-1996*, Journal of Economic Inequality, v. 4, 1, pp. 1-32, **TD No. 487 (March 2004)**.

M. BOFONDI and G. GOBBI, *Information barriers to entry into credit markets,* Review of Finance, Vol. 10 (1), pp. 39-67, **TD No. 509 (July 2004).**

LIPPI F. and W. FUCHS, *Monetary union with voluntary participation,* Review of Economic Studies, 73, pp. 437-457 **TD No. 512 (July 2004).**

GAIOTTI E. and A. SECCHI, *Is there a cost channel of monetary transmission? An investigation into the pricing behaviour of 2000 firms*, Journal of Money, Credit, and Banking, v. 38, 8, pp. 2013-2038 **TD No. 525 (December 2004)**.

A. BRANDOLINI, P. CIPOLLONE and E. VIVIANO, *Does the ILO definition capture all unemployment?,* Journal of the European Economic Association, v. 4, 1, pp. 153-179, **TD No. 529 (December 2004)**.

A. BRANDOLINI, L. CANNARI, G. D'ALESSIO and I. FAIELLA, *Household Wealth Distribution in Italy in the 1990s,* In E. N. Wolff (ed.) International Perspectives on Household Wealth, Cheltenham, Edward Elgar, **TD No. 530 (December 2004)**.

A. NOBILI, *Assessing the predictive power of financial spreads in the euro area: does parameters instability matter?,* Empirical Economics, v. 31, 4, pp. , **TD No. 544 (February 2005)**.

L. GUISO and M. PAIELLA, *The Role of Risk Aversion in Predicting Individual Behavior, In P. A. Chiappori e C. Gollier (eds.) Competitive Failures in Insurance Markets: Theory and Policy Implications, Monaco, CESifo*, **TD No. 546 (February 2005).**

G. M. TOMAT, *Prices product differentiation and quality measurement: A comparison between hedonic and matched model methods*, Research in Economics, No. 60, pp. 54-68, **TD No. 547 (February 2005)**.

M. CARUSO, *Stock market fluctuations and money demand in Italy, 1913 - 2003*, Economic Notes, v. 35, 1, pp. 1-47, **TD No. 576 (February 2006)**.

R. BRONZINI and G. DE BLASIO, *Evaluating the impact of investment incentives: The case of Italy's Law 488/92.* Journal of Urban Economics, vol. 60, n. 2, pag. 327-349, **TD No. 582 (March 2006).**

A. DI CESARE, *Do market-based indicators anticipate rating agencies? Evidence for international banks*, Economic Notes, v. 35, pp. 121-150, **TD No. 593 (May 2006).**

FORTHCOMING

S. MAGRI, *Italian Households' Debt: The Participation to the Debt market and the Size of the Loan*, Empirical Economics, **TD No. 454 (October 2002)**.

LIPPI F. and S. NERI, *Information variables for monetary policy in a small structural model of the euro area*, Journal of Monetary Economics **TD No. 511 (July 2004)**.

DEDOLA L. and S. NERI, *What does a technology shock do? A VAR analysis with model-based sign restrictions*, Journal of Monetary Economics **TD No. 607 (December 2006).**