Laboratory of Data Analysis
University of Jyväskylä

EUREDIT - WP4.5, WP5.5 Internal reports

WP4.5 & WP5.5: The coding of TS-SOM experiments with Euredit evaluation data sets.

Pasi P. Koikkalainen - University of Jyväskylä
Ismo Horppu - - University of Jyväskylä

Contents

Summary

This report describes how we have applied the tree-structured self-organizing map (TS-SOM) for statistical editing and imputation in the Euredit project. We expect that the reader is already familiar about the basic TS-SOM methodology, which is described in reports D5.5.1 and D5.4.1. In addition to previous documents, this report presents also the practicalities of the corresponding editing and imputation procedures.

Our methodology has been tested using all development data sets, provided by the Euredit project, excluding the "Insiders data", which will be tested later. When available, both Y2 (missing data) and Y3 (missing and erroneous) have been used in the experiments.

In theory all our editing and imputing algorithms can be used automatically without user interaction. In practice, however, the methodology is at its best when used with other data analysis tools that support interative and interactive development. This includes an appropriate selection of background information, the coding of varibles, as well as the values of the training parameters of the TS-SOM algorithm. These inputs are sometimes essential in order to obtain good results in statistical editing and imputation.

To minimize the role of user choices we have applied a simple automated trial-and-error type of paramer search with some of the data sets. The algorithm makes several randomly chosen editing and imputation experiments with different parameter selections.

The procedures that gave the best editing and imputation results by the end of July 2002 are briefly outlined in this paper. Due the large number of possible choices we have coded our experiments with letters and numbers, $A0, B1, \ldots, L4$, which define the parameters of the experiment. In addition for each edited and imputed variable the background information (covariates) is often selected differently, depending on the type of experiment.

It should be noted that the SOM methodology is still under development. In the future we expect improvements both in the usability of our software and in the editing & imputation performance of the algorithms.

1 Introduction

One should recall from our previous reports that the TS-SOM is similar to principal curves (nonlinear principal subspaces) and some clustering algorithms. The basic idea of how to use TS-SOM for editing and imputation is to build a SOM representation, a lower dimensional regression surface of data, and project all data samples (records) onto this model. The projection point on the TS-SOM surface, where the data record is projected onto represents the expectation (actually sample mean) of similar samples.

For editing and imputation purposes we estimate also the deviations between the SOM model and the observed training samples. We may suspect that the sample is outlier if its distance from the projection point on the SOM surface is radically larger than the deviation of other samples from that point.

In the case of imputation the projection between the sample and the SOM surface can be defined for incomplete samples also. Using the SOM model, the observed part of the sample defines a conditional distribution of unobserved values. We may then apply any standard imputation procedure.

We also remind the reader that in the SOM algorithm the principal surface is constructed with a set of nodes, which can be understod as data clusters. Therefore all computations are actually done with clusters of data.

1.1 Editing and Imputation strategies

Our editing and imputation methodology is based on three steps. First the TS-SOM is used as an overall model of the distribution. Then all samples are compared to the TS-SOM model to detect outliers, which are marked as potential errors for later processing. Finally the missing values are imputed by selecting samples from or with help of the TS-SOM model. On all levels there are several issues that are specific to the handling of erroneous and missing data.

In modelling the role of TS-SOM algorithm is to model the true, but unknown, multivariate distribution from data. The challenge of this step is the handling of incompete and erroneous observations during modelling. Therefore we have developed new robust training algorithms that can handle both missing variables and data outliers during the training.

In editing the task is to investigate if sample is well expained by the TS-SOM model. If not, it could be an error.

In imputation the concern is to select as good values as possible for the missing ones according to our approximative TS-SOM model. First the observed part of an incomplete sample is used to conditionalize the model. Then the actual values are selected by using an imputation stategy, which is actually a local imputation model, conditionalied by the TS-SOM and observed values of the incompete sample.

Except some implementation details and errors, which are now corrected and described in this paper, the editing and imputation methodologies are explained in our previous reports of workpagages WP4.5 (D4.5.1) and WP5.5 (D5.5.1).

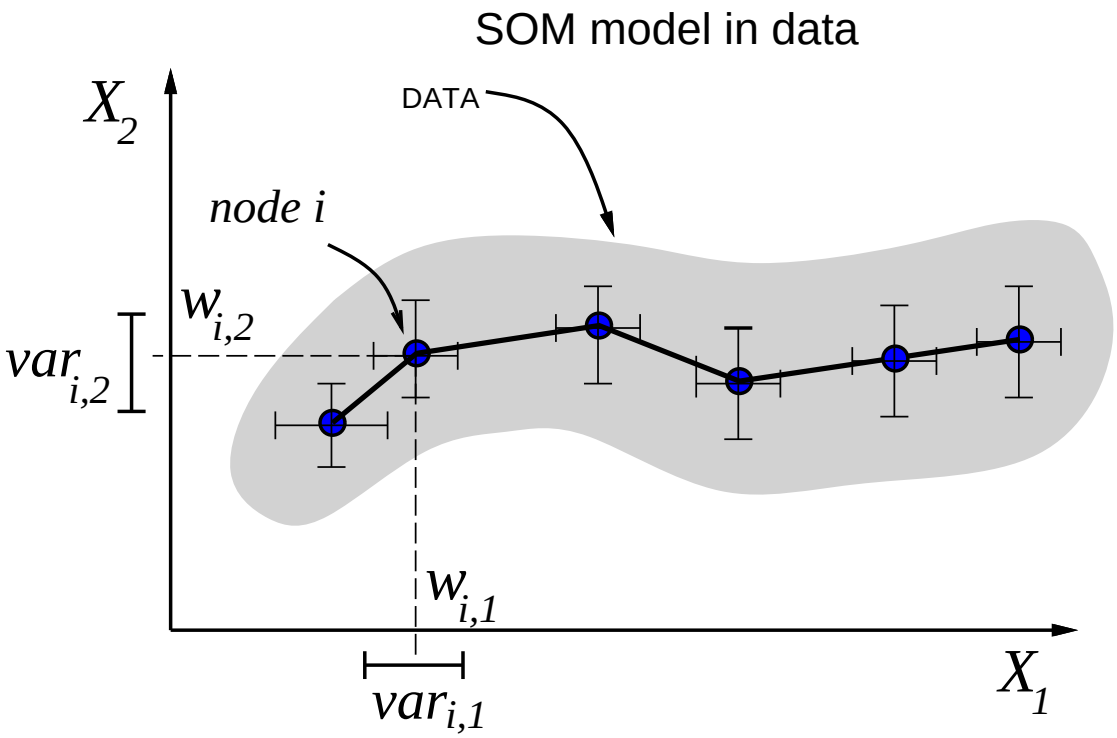
1.2 Modelling: an improved update rule

The purpose of the SOM training algorithm is to find SOM weight parameters $w_{i,r}$. These weights are cluster means that approximate the expectation $w_{i,r} \approx \mathbb{E}[X_r | b(\mathbf{X}) = i]$ of the discretized implementation of a principal curve (see D5.5.1 for more details). In other words, the weights define how the SOM is located in the data.

When the SOM algorithm is used for imputation, more information than just simple weight values is often required. For example, local variances or other higher moments would allow us to model data distribution around the SOM regression surface. This additional information would then help

us to improve the distributional properties of imputed values. The problem is that higher moments are costly to compute in multivariate cases. Therefore we have done a compromise by assuming a diagonal covariance matrix for every SOM node, which also illustrated in Fig 1. Now variances, or actually standard deviations, need to be computed only for each of the variables (data axes).

Chart 1: Node mean and variance parameters of the self-organizing map.



In addition to variances the most notable change to previous reports is the way how the update step of algorithms 3 (D5.5.1/imputation) and 4 (D4.5.1/edit) is done. Rather than replacing the missing values with node mean, the value of SOM node weight parameter is computed directly from observed ones. The missing values are simply omitted, which is also noted by omitting their influence to the node priors.

The corrected TS-SOM training algorithm for one layer, with new update rule (setp 2.3) is shown in Algorithm 5.

Algorithm 5: New estimation procedure for SOM parameters.

2. Train layer: l

Step 2.1 *Use lookup search i*

$\Omega_i = \{j \mid i \in \mathrm{b}_N^l(\ \mathbf{X}^{\mathrm{obs}}(j)\ )\ \};$

Step 2.2 *For every node i*

Step 2.2.1 (*Compute node priors for every variable r*)

$N_{i,r}^{sw,\mathrm{obs}} = \sum_{j \in \mathtt{Robs}} s\hat{w}(j), \quad \text{where } \mathtt{Robs} = \{j \mid I_r^{\mathrm{mis}}(j) = \mathtt{false}, j \in \Omega_i\}$

Step 2.2.2 (*Compute node means for every variable r*)

$\overline{x}_{k,r}^{\mathrm{obs}} = \frac{1}{N_{k,r}^{s\hat{w},\mathrm{obs}}} \sum_{j \in \Omega_k, I_r^{\mathrm{mis}}(j) = \mathtt{false}} s\hat{w}(j) X_r^{\mathrm{obs}}(j)$

Step 2.2.3 (*Compute node deviations for every variable r*)

$\sigma*_{k,r}^{\mathrm{obs}} = \frac{1}{N_{k,r}^{s\hat{w},\mathrm{obs}} - 1} \sum_{j \in \Omega_k, I_r^{\mathrm{mis}}(j) = \mathtt{false}} s\hat{w}(j) |\overline{x}_{k,r}^{\mathrm{obs}} - X_r^{\mathrm{obs}}(j)|$

Step 2.3 *For every node i*

Step 2.3.1 *Compute node positions for every variable r*

$w_{i,r}^{t+1} = \frac{1}{\sum_k h_{i,k} N_{k,r}^{s\hat{w},\mathrm{obs}}} \sum_k h_{i,k} N_{k,r}^{s\hat{w},\mathrm{obs}} \overline{x}_{k,r}^{\mathrm{obs}}$

Step 2.3.2 *Compute smoothed deviations for every variable r*

$\sigma_{k,r}^{t+1} = \frac{1}{\sum_k h_{i,k} N_{k,r}^{s\hat{w},\mathrm{obs}}} \sum_k h_{i,k} N_{k,r}^{s\hat{w},\mathrm{obs}} \sigma*_{k,r}^{\mathrm{obs}}$

2.4. Repeat layer training until converged:
If $\|\mathbf{W}^{new} - \mathbf{W}^{old}\| > \delta$ GOTO 2.

Here $\mathbf{X}^{\mathrm{obs}}(j)$ is the observed part of sample j, Ω_i is the set of best matching samples that are associated to TS-SOM node i, $s\hat{w}(j)$ is the sampleing weight of sample j, and $w_{i,r}$ is the r^{th} weight value of node i.

1.3 Modelling: robust training options

The current implementation of robust training can be done in two ways. Both methods are based on outlier detection that was descibed in report D4.5.1.

The first method is the same that was in the earlier report. Samples that are considered to be outliers are simply omitted. This is done by marking those values as missing:

$$I_r^{\mathrm{mis}}(j) = \begin{cases} \mathtt{true} & \text{if } |w_{i,r} - X_r(j)| > \theta_1 \sigma_{i,r} \\ \mathtt{true} & \text{if } I_r^{\mathrm{mis}}(j) = \mathtt{true} \\ \mathtt{false} & \text{otherwise} \end{cases}$$

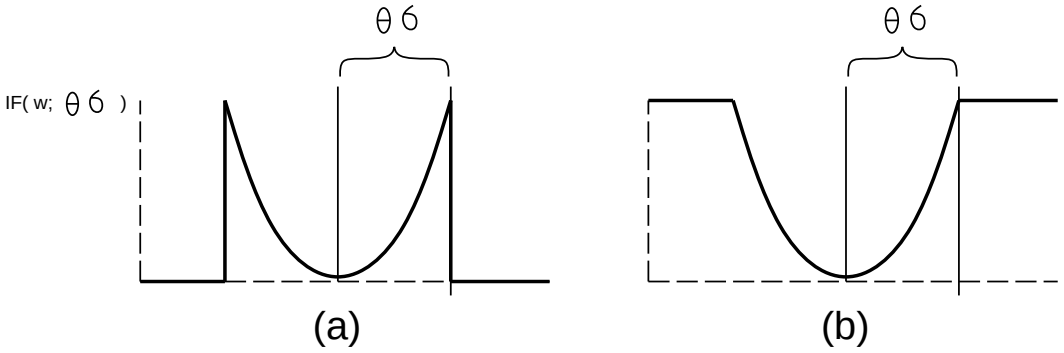
Our second robust training option is to use Huber M-estimator. Here no observed samples are omitted

but possible outliers are moved closer to sample mean by replacing the original sample value X_r with X'_r such that

$$X'^{\text{obs}}_r(j) = \begin{cases} X^{\text{obs}}_r(j) & \text{if } |w_{i,r} - X^{\text{obs}}_r(j)| < \theta_1 \sigma_{i,r} \\ w_{i,r} - \theta_1 \sigma_{i,r} & \text{if } w_{i,r} - X^{\text{obs}}_r(j) > 0 \\ w_{i,r} + \theta_1 \sigma_{i,r} & \text{otherwise.} \end{cases}$$

In both robust training options the influence of errors is controlled by parameter θ_1. For example, value $\theta_1 = 3$ indicates that we allow sample X_r to differ at maximum three times standard deviatiation $\sigma_{i,r}$ from the node mean $w_{i,r}$ of the best matching SOM node i. This effect on influence functions is also illustrated in Fig. 2.

Chart 2: Influence functions for two robust estimators, used in SOM training: (a) omitting outliers and (b) Huber estimator.



Influence is same for both estimated parameters: node means $w_{i,r}$ and deviations $\sigma_{i,r}$. A

1.4 Robust outlier detection

After training the TS-SOM with robust options, it is relatively easy to imagine how to use a SOM layer for outlier detection. First for each observation $\mathbf{X}(j)$ the best matching unit $i = b^l(\mathbf{X}(j))$ is searched. Then for each variable we examine its distance from the mean and compare this to some predefined distance, which should be defined as a function of error probability.

It is more difficult, however, to decide the probability of an error, since there is no objective criteria for this. In lack of any good theory we decided to approximate error probabilities with a simple outlier to *Normal pfd* assumption as follows.

In the case of **continuous variables** $X_r \in \mathbb{R}$ a distance P_e is computed first:

$$P_e = 1 - \frac{1}{Z} \exp\left(-\frac{\|X_r - w_{i,r}\|^2}{2 * \sigma^2_{i,r} * \theta_2}, \right)$$

where $Z = \sqrt{2\pi\sigma^2_{i,r}\theta_2}$. Then the final probablility of an error is select to be

$$\Pr(\texttt{error}|X_r) = \begin{cases} P_e & \text{if } P^{\text{cut}}_r < P_e \leq 1 \\ 1 & \text{if } P_e > 1 \\ 0 & \text{if } P_e < P^{\text{cut}}_r, \end{cases}$$

where P^{cut}_r is "cut probability" parameter for variable r. The parameter θ_2 has now less significant role than in our previous report (D5.4.1). It reshapes the normal assumption of correct data, if nessessary. In most cases we can ignore it (set $\theta_2 = 1$).

In case of **categorial variables** $X_r \in \{0, 1\}$ even simpler outlier detection mechanism was used. Now

$$P_e = |w_{i,r} - X_r|$$

and

$$\Pr(\texttt{error}|X_r) = \begin{cases} P_e & \text{if } P_e > P_r^{\text{cut}} \\ 0 & \text{if } P_e < P_r^{\text{cut}}, \end{cases}$$

.

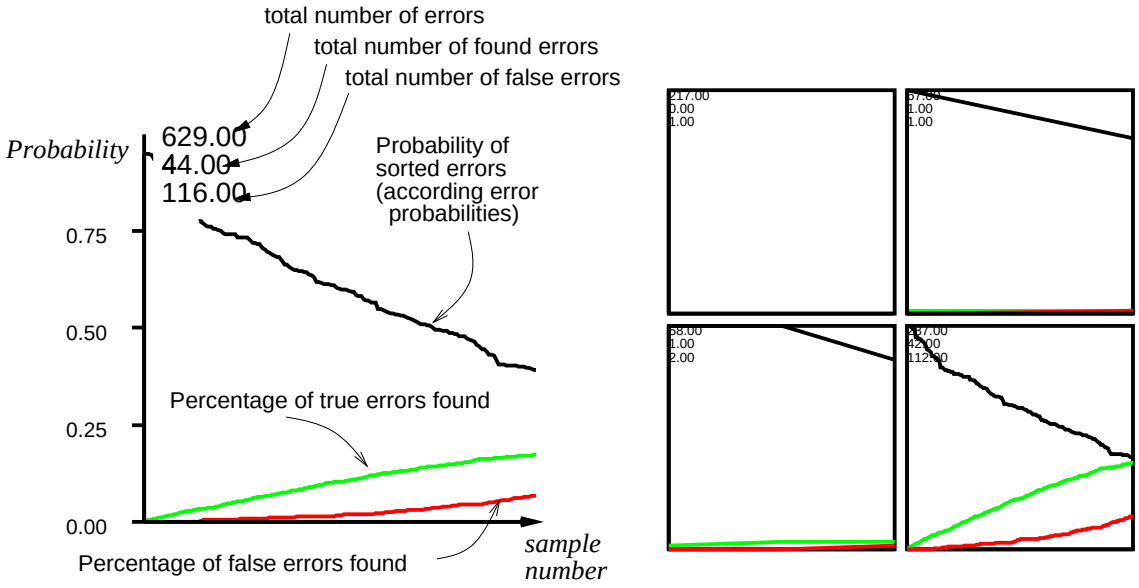
Because it is quite difficult to say what is the relation with our error probability $\Pr(\texttt{error}|X_r)$ to true errors of data, the P_r^{cut} thresholds were computed by assuming the expected number of errors from data sets. Using development data we had an idea of the per centage of errors p_r^{errors} of the variable r. The cut probability was then selected by softing errors such that

$$\Pr(\texttt{error}|X_r(a_1)) > \ldots \Pr(\texttt{error}|X_r(a_n)) > P_r^{\text{cut}} > \Pr(\texttt{error}|X_r(a_{n+1})) > \ldots > \Pr(\texttt{error}|X_r(a_N)).$$

where N is the number of observations and $p_r^{\text{errors}} = n/N$.

The error detection mechanism can also be visualized as depicted in Fig. 3 when true data is available. These figures give an impression, based on development data sets, how well we may exepect our error detection to work. When errors are sorted according to their probabilities the graphs from top to bottom illustrate: error probablilities of samples, cumulation of found errors, cumulation of false errors. Numbers on the top left of the node denote the total number of errors associated to this node, number of found errors up to P_r^{cut} and number of false errors up to P_r^{cut}.

Chart 3: Visualization of found true and false errors from the development data set with SOM



One SOM node **Four nodes in SOM layer 1**

2 Imputation procedures

We have implemented six different imputation procedures. All imputation routines use the TS-SOM as a model of "correct" distribution. First step of the imputation is to find a set of `Nb` best candiatate

neurons (clusters)

$$b_N(\mathbf{X}^{obs}) = \{k_1, k_2, \ldots, k_{\mathtt{Nb}} \mid d_{k_1}(\mathbf{X}^{obs}) < d_{k_2}(\mathbf{X}^{obs}) < \ldots < d_{k_{\mathtt{Nb}}}(\mathbf{X}^{obs})\},$$

for the observed part of sample $\mathbf{X}^{\mathrm{obs}}$. Imputation is done using one neuron only, which can be selected either deterministically (closest neuron):

$$\texttt{best node is } i = k_1,$$

or via random selection

$$\texttt{best node is } i = k_{\mathtt{RAND}(1..N_b)}$$

Actual values are chosen according to local neuron statistics, where the six possibilities for missing X_r^{mis} values are:

0 use mean values
$X_r^{\mathrm{imp}} = w_{i,r}$

1 pick a random sample from truncated Normal pdf.
$X_r^{\mathrm{imp}} \sim N(w_{i,r}, \sigma_{i,r}^2)_{|w_{i,r}-2\sigma_{i,r}<X_r^{\mathrm{imp}}<w_{i,r}+2\sigma_{i,r}}$

2 pick a random sample according to uniform pdf.
$X_r^{\mathrm{imp}} \sim U(w_{i,r} - \sigma_{i,r}, w_{i,r} + \sigma_{i,r})$

3 Use random donor from cluster i
$X_r^{\mathrm{imp}} = X_j$, where $j = Rand(0..N_i), j \in \Omega_i$

4 Use nearest neighbor donor from cluster i
$X_r^{\mathrm{imp}} = X_r(k)^{\mathrm{obs}}$, where $k = \arg\min_{k'} \|\mathbf{X}(k')^{\mathrm{obs}} - \mathbf{X}(j)^{\mathrm{obs}}\|$

5 Use node specific MLP regression model for imputation
$X_r^{\mathrm{imp}} = f_r^{\mathrm{MLP}_i}(\mathbf{X}^{\mathrm{obs}}|i)$

3 TS-SOM editing and imputation in six steps.

Editing and imputation procedures of the TS-SOM are implemented under six operational steps in our NDA software. The flowchart of these operations is depicted in Fig. 4. First step i) is data preprocessing, ii) is the training of the TS-SOM neural network, iii) is the estimation of cluster statistics, iv) is the projection of (incomplete) samples onto the TS-SOM model, v) is the editing and vi) is the imutation step.

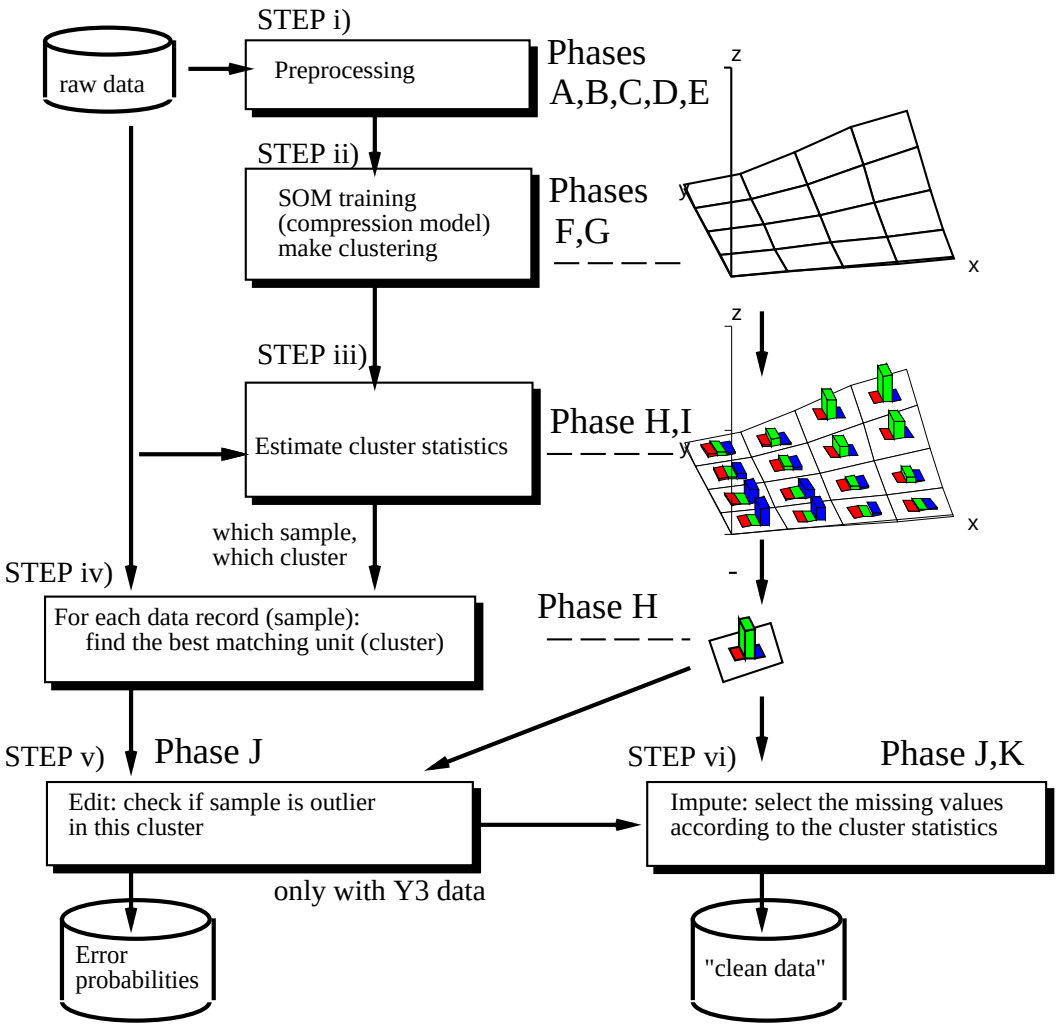
Excluding the problem of variable selection (covariates) variations of the use of the TS-SOM for editing and imputation can be summarized as:

i) Data preprocessing includes scalar and vector level equalization of variables, possible transformations like the log-transform and rule based editing of obvious errors. In our coding these are noted as phases A,B,C, and E, as described later.

ii) Training step of the TS-SOM neural network. It can be done either with normal TS-SOM training, which assumes fully observed clean data, or with our new robust incomplete data version of the TS-SOM algorithm. This step makes a clustering of all data records, where the number of clusters is defined by the TS-SOM layer L. The other training options of the TS-SOM denoted with letters F and G.

iii) Computation of cluster statistics is done by estimating the mean and variance of variables, which are the ones we are editing or/and imputing. Estimates are computed for each of the clusters. There are several ways to do this. In the simplest case these are the "weight" parameters of the TS-SOM, computed in step ii), and no additional work needs to be done. Some preprocessing

Chart 4: Operations of the TS-SOM based editing and imputation software that have been used in our experiments.



operations, however, prevent us using TS-SOM weights are statistical estimates, and then the statistics must be computed from original raw data (external data). In the following chapters we denote the computation of cluster statistics with letter I.

iv) Projection of data samples onto the TS-SOM surface (node, cluster) is done during the training and also in the post processing, in editing and imputation steps. In the simplest case the sample is projected always onto the closest, most similar TS-SOM node. Another possibility, which should be used with missing data, is to select some small number of canditates and then select the projection point in random. This is our option letter H.

v) Currently we have support for one editing procedure only. This is a simple outlier detection routine that assigns a probability for each variable of a sample that is radically different than other samples in the cluster. In practice this is done such that in each cluster for categorial variables we estimate the binomial distribution and for continuous variables we estimate Normal pdf to which the samples are compared to.

vi) There are currently six dirrerent imputation routines for incomplete observations which are assigned to some TS-SOM cluster. We may use the cluster mean values directly, select the

missing values from estimated cluster specific pdf (Uniform or Normal), use a donor from the other samples of the cluster, or build a MLP type of regression model inside the cluster. These options are noted with letter K. In addition we may choose to impute outliers also, which is controlled by option letter J.

3.1 The coding of operational options

There are a total of twelve different operational options, A,B,C,D,E,F,G,H,I,J,K and L, for TS-SOM editing and imputation procedures. Each of the options takes two or more values that determine how the actual editing and/or imputation is done. These options are summarized in Fig. 5. In practice only a subset of option combinations is needed, depending on the type of data and problem.

Chart 5: Options for TS-SOM editing and imputation procedures.

A	**Preprocessing option 1**	**A0** Check of bounds=NO	**A1** Check of bounds=YES				
B	**Preprocessing option 2**	**B0** Log−transform=NO	**B1** Log−transform=YES				
C	**Preprocessing option 3**	**C0** Min−max equalization=NO	**C1** Min−max equalization=YES				
D	**Preprocessing option 4**	**D0** Vector length equaliztion=NO	**D1** Vector length equaliztion=YES				
E	**Preprocessing option 5**	**E0** Edit rules =NO All edit rules ingored	**E1** Simple bounds check	**E2** Rule based editing used before SOM training			
F	**Data compression option 1**	**F0** Using incomplete data training (our new IC−SOM algorithm)	**F1** Using only fully observed records with SOM (normal SOM)				
G	**Data compression option 2**	**G0** Basic SOM (no robustness in training)	**G1** Discard outliers during SOM−training	**G2** Using robust (Huber estimator) with SOM−training			
H	**Selection of best matching unit**	**H0** Probabilistic	**H1** Deterministic				
I	**Compute imputation statistics from**	**I0** SOM training data	**I1** External data				
J	**Outlier handling**	**J0** Mark as missing (do not impute)	**J1** Impute most obvious outliers only	**J2** Impute all outliers			
K	**Impute missing value within cluster from**	**K0** Cluster mean	**K1** Normal pdf	**K2** Uniform pdf	**K3** Random donor	**K4** Nearest neighbor donor	**K5** MLP (regression)
L	**TS−SOM layer**	**L0** 1−cluster	**L1** 4−clusters	**L2** 16−clusters	• • •		**L6** 4096−clusters

The first five binary options A,B,C,D and E control the preprocessing of data before the TS-SOM training.

A (bounds check) is typically on (option A1). It checks if the variable is between an acceptable range

of values. If bounds chech fails the variable is marked as missing.

B (log-transform) is usually used (B1) with ABI and EPE data sets. Operation is automatically inverted in post processing.

C (min-max equalization) is typically on (C1). This converts the (robust) variation of samples between 0 and 1 for each of the samples. Operation is automatically inverted in the post processing.

D (vector lenght equalization, normalization). This operation equalizes the norm of the input sample $\mathbf{x}$ to unity: $\|\mathbf{x}\| = 1$. Operation is not invertible and thus external/original data must be used instead of SOM weights in step iii) (estimation of cluster statistics).

E (edit rules), if available, can be used (E1) to correct obvious hard errors before SOM training. In some cases this improves the imputation results significantly.

Options F and G control the TS-SOM training algorithm.

F (incomplete data training), option is mostly on (F0) in the current experiments. This allows all data to be used in the modelling step, while option F1 rejects the incompete samples.

G (robustness) options (G1) or (G2) are mostly used when modelling Y3 type of erroneous data. This option tries to omit outliers during the training step.

Option H has effect only for incomplete and suspected outlier samples.

H (Projection operation) is usually probabilistic (H0), which adds randomness to the selection of the best matching unit (node, cluster) for the observed incompete sample. Option H1 selects always strictly closest node.

Option I is typically consequense of preprocessing options that prevent us from using SOM weights for imputation.

I (cluster statistics), where value (I0) indicated that TS-SOM weights are used for imputation. Otherwise the statistics for clusters must be estimated from original data (I1). Note that option (I1) allows us to use different variables for model building and actual imputation.

Option J defines what will be done with the outliers. This option is used only with Y3 data sets.

J (Oulier handling). It is not advisible to impute all outliers. Option (J0) omits all outlier corrections. Outliers are only marked as missing data. Option (J1) imputes most severe (probable) outliers and with option (J2) all detecte outliers are imputed.

Option K selects the imputation procedure for incomplete observations. All procedures are applied after the projection operation that selects the "best" cluster for a given incomplete sample.

K0 Imputation procedure (K0) completes the missing values from cluster mean. When the number of cluster is very large this is almost same as the nearest neighbor donor imputation.

K1 Assumed that data inside the cluster is Normal distributed and selects the value randomly from this pdf.

K2 Assumed that data inside the cluster is Uniform distributed and selects the value randomly from this pdf.

K3 Selects random donor from the cluster of samples.

K4 Finds the nearest neighbor donor from the cluster.

K5 In this option a MLP neural network has been trained using the samples of the TS-SOM cluster. This sub-model is then used as a predictor of the sample values.

4 Configuring TS-SOM edit using UK ABI development datasets

We give an example of detecting editing parameters for major variables which are TURNOVER, EMPTOTC, PURTOT, TAXTOT, ASSACQ and ASSDISP. Development UK ABI sector 1/97 Y3 and true datasets were used to detect the editing parameters. The parameters for a continuous variable are sigma1, sigma2 and cut probability. We used TS-SOM layer 2 (= 16 neurons/clusters) for editing.

First of all it is very important to log transform exponential scale variables. The reason is that TS-SOM editing uses normal distribution, for continuous variables, to detect errors. Secondly we use robust min-max equalization using 5% and 95% fractiles. The explanatory variables are TURNREG and EMPREG which are registered turnover and employment size group from the register. Both of the variables have not been perturbed. In addition TAXRATES variable is used with TAXTOT variable.

We present two experiments which are conservative(1) and aggressive(2) (when considering error detection). The aim of the the latter one is to detect more false positives, and not much more false negatives, than the first one. The latter one is derived from the first one by modifying sigma1 and sigma2 values.

The edit parameters for variables and experiments are shown in tables 1 and 2.

Variable	**Sigma1/Sigma2** (expr. 1)	**Sigma1/Sigma2** (expr. 2)
TURNOVER	3.25/2.15	3.25/1.85
EMPTOTC	3.25/2.15	3.25/1.85
PURTOT	3.25/2.15	3.25/1.85
TAXTOT	4.0/1.0	3.75/0.7
ASSACQ	3.5/1.15	3.5/1.145
ASSDISP	3.5/1.15	3.5/1.145

Table 1: Edit sigma1 and sigma2 parameters

Variable	**Edit cut probability** (expr. 1 and expr. 2)
TURNOVER	0.45
EMPTOTC	0.4
PURTOT	0.4
TAXTOT	0.5
ASSACQ	0.2
ASSDISP	0.2

Table 2: Edit cut probability parameters

Table 3 shows the error detection statistics. In the table N is total errors in true data, N(TRUE) is false positives, N(FALSE) is false negatives.

Figures 1-12 have 3 graphs which correspond TS-SOM layers 0-2 (1, 4 and 16 neurons/clusters). The figures show the TS-SOM error probability curves (black/darkest linegraph) which start from highest error probability and end to lowest. In addition they include cumulative error functions for false positives (green - the lightest linegraph) and false negatives (red - the second darkest linegraph). The X-axis is arranged according to error probabilities of records from highest to lowest, however one must realize that the edit cut probability cuts the amount of error probabilities of records in the linegraphs. The Y-axis in the graphs is the error probability for TS-SOM error probability curve (the darkest linegraph). For the two other linegraphs the Y-axis is cumulative error probability. The Y-axis is labelled only in the layer 0. Upper left corner of the graphs have field values (from top to bottom): number of true data errors, number of false negatives and number of fales positives. Empty graphs

mean that TS-SOM has not detected any false positives (or false negatives).

Variable	**Expr.**	**N**	**N(TRUE)**	**N(FALSE)**
TURNOVER	1	241	175	47
	2	241	189	60
EMPTOTC	1	332	111	25
	2	332	136	29
PURTOT	1	629	116	44
	2	629	130	55
TAXTOT	1	482	201	56
	2	482	292	276
ASSACQ	1	248	58	55
	2	248	43	79
ASSDISP	1	224	49	23
	2	224	57	44

Table 3: ABI Y3 error detection statistics

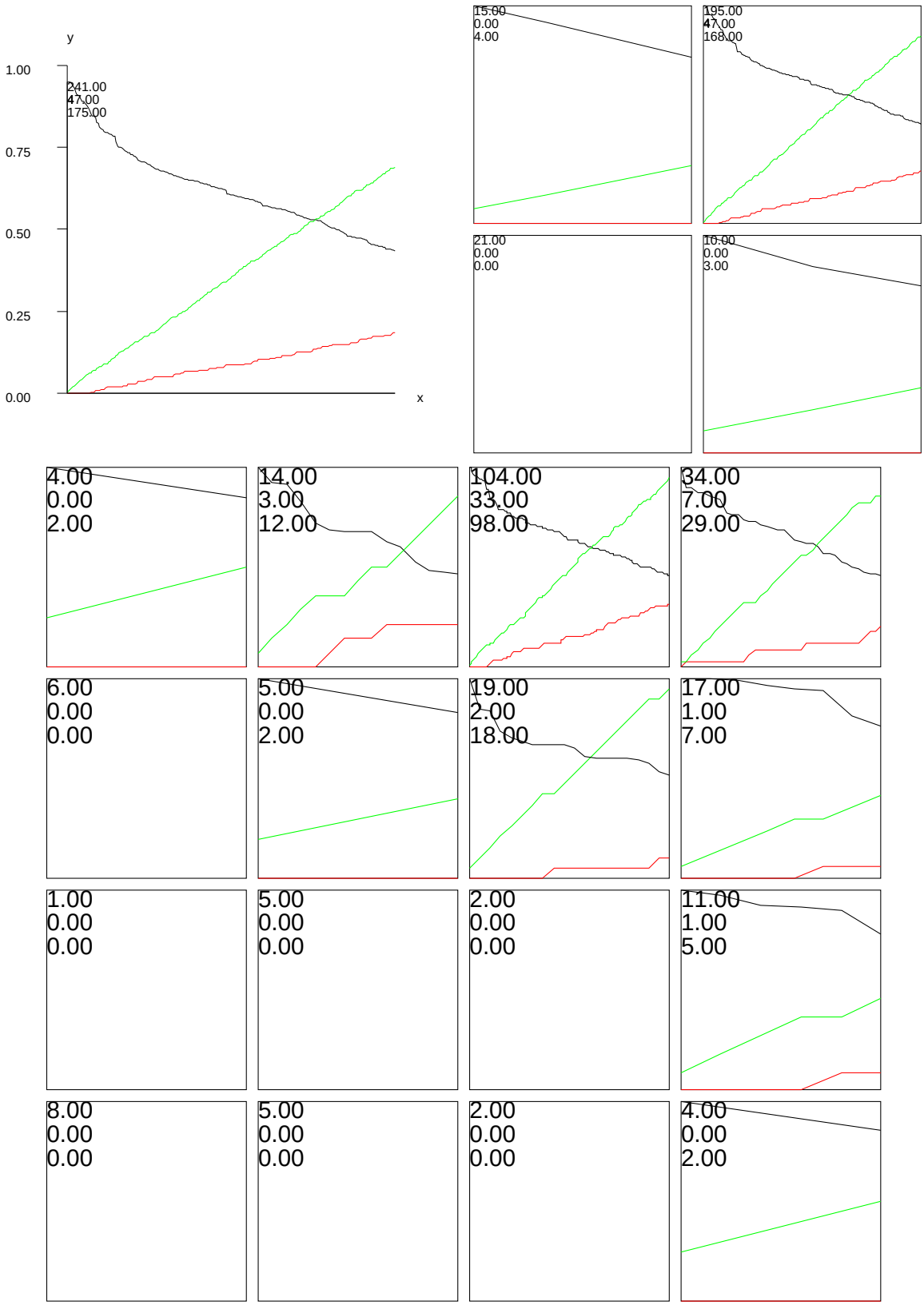


Fig. 1: Error detection graphs for TURNOVER (expr. 1)

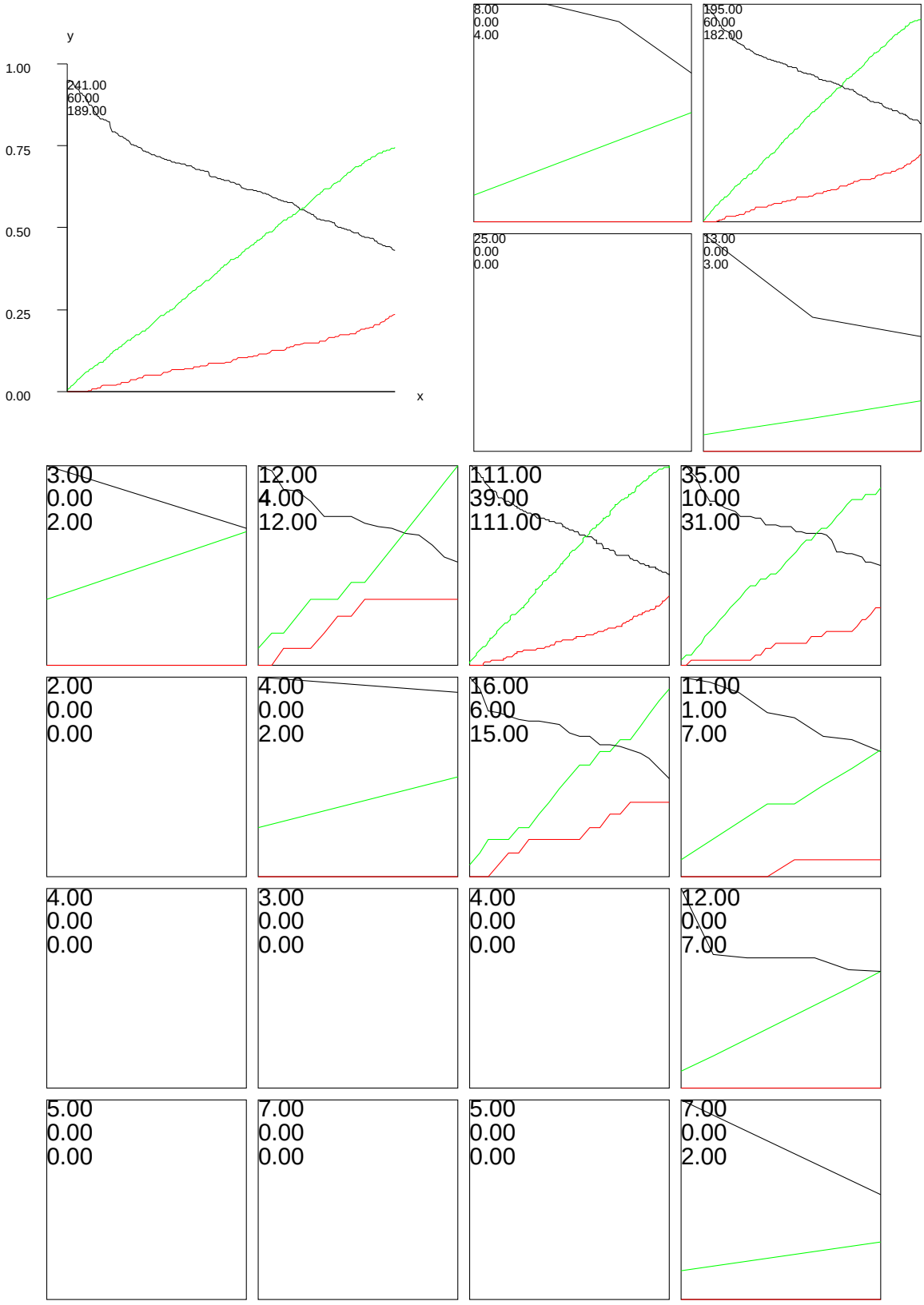


Fig. 2: Error detection graphs for TURNOVER (expr. 2)

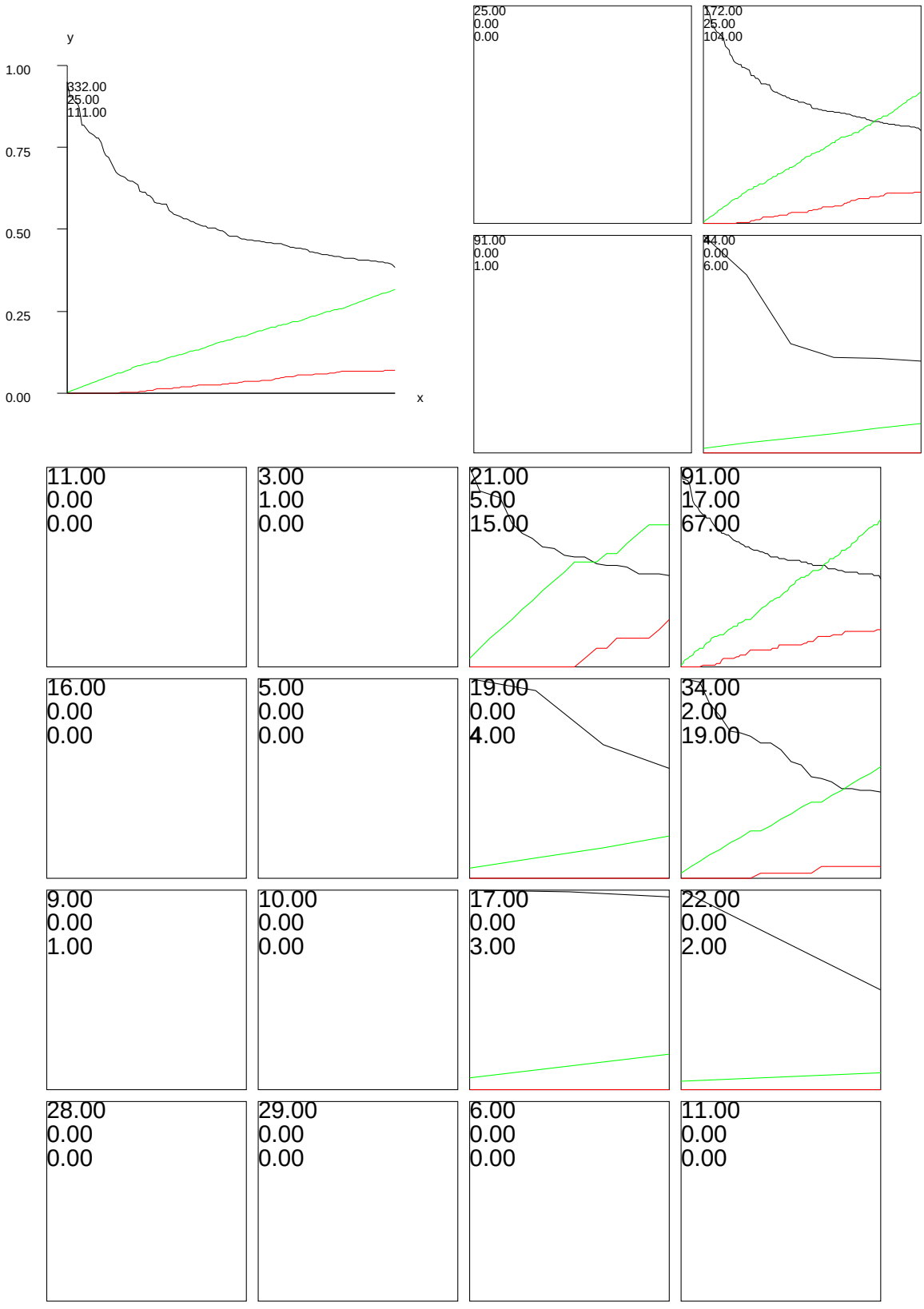


Fig. 3: Error detection graphs for EMPTOTC (expr. 1)

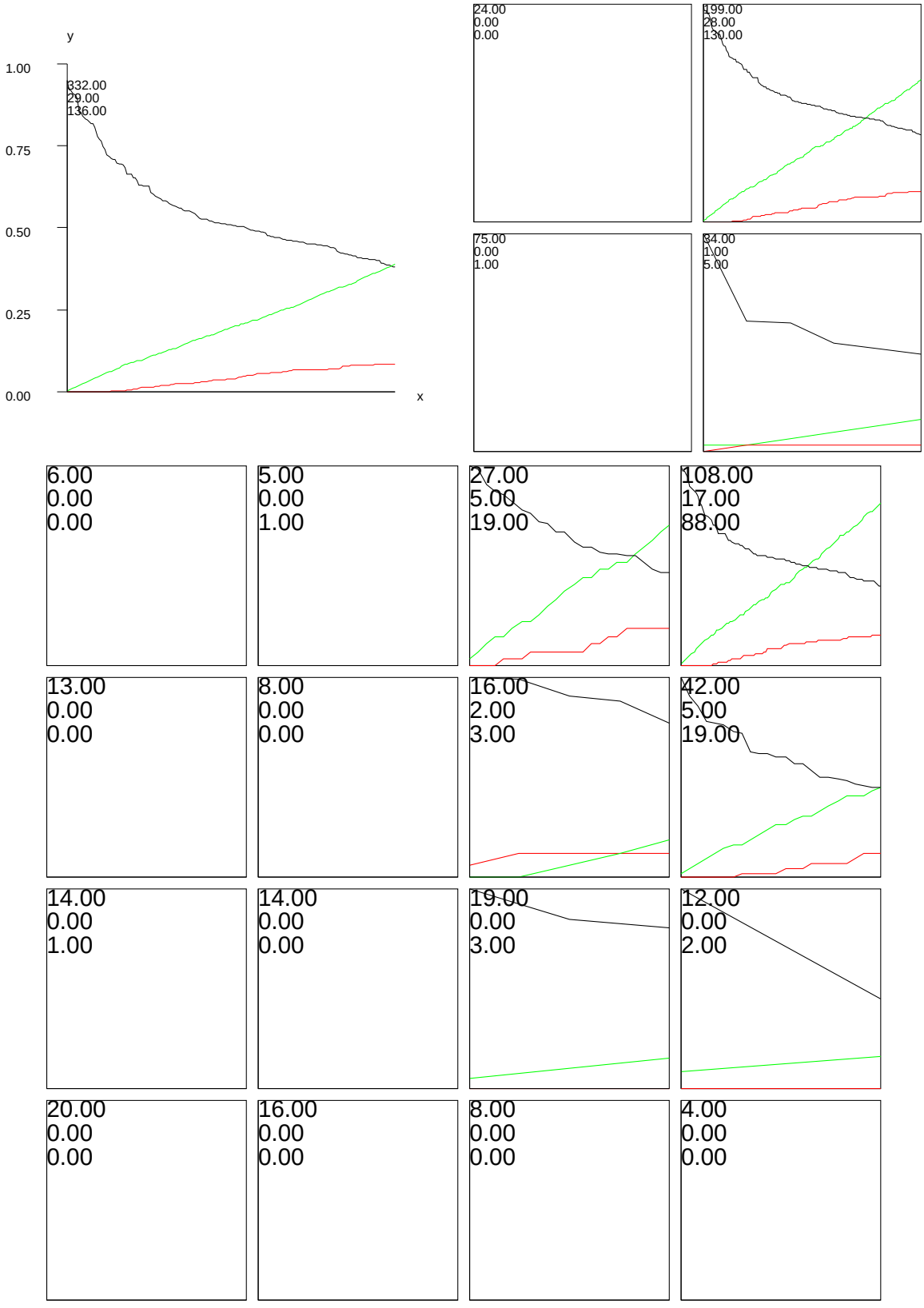


Fig. 4: Error detection graphs for EMPTOTC (expr. 2)

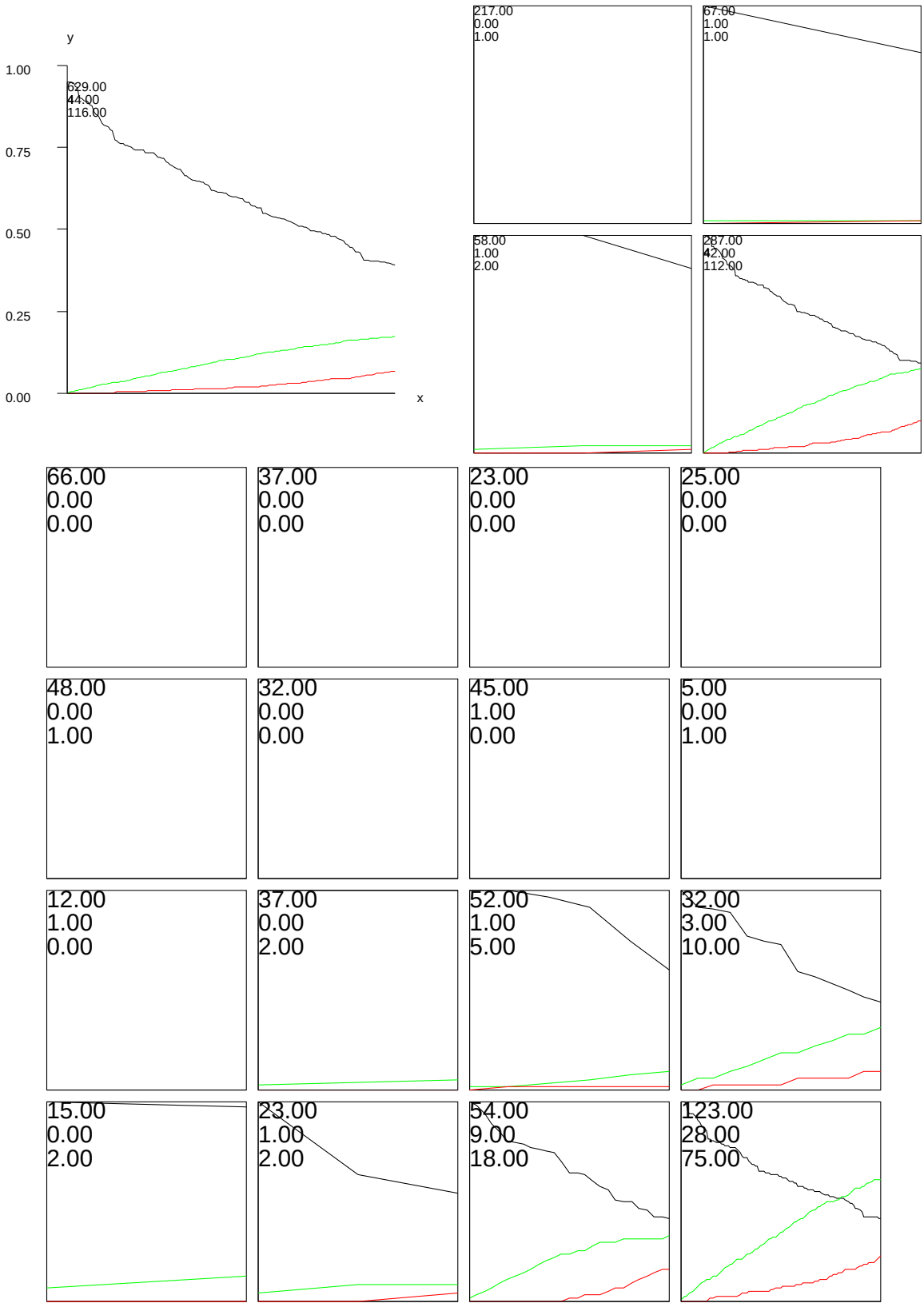


Fig. 5: Error detection graphs for PURTOT (expr. 1)

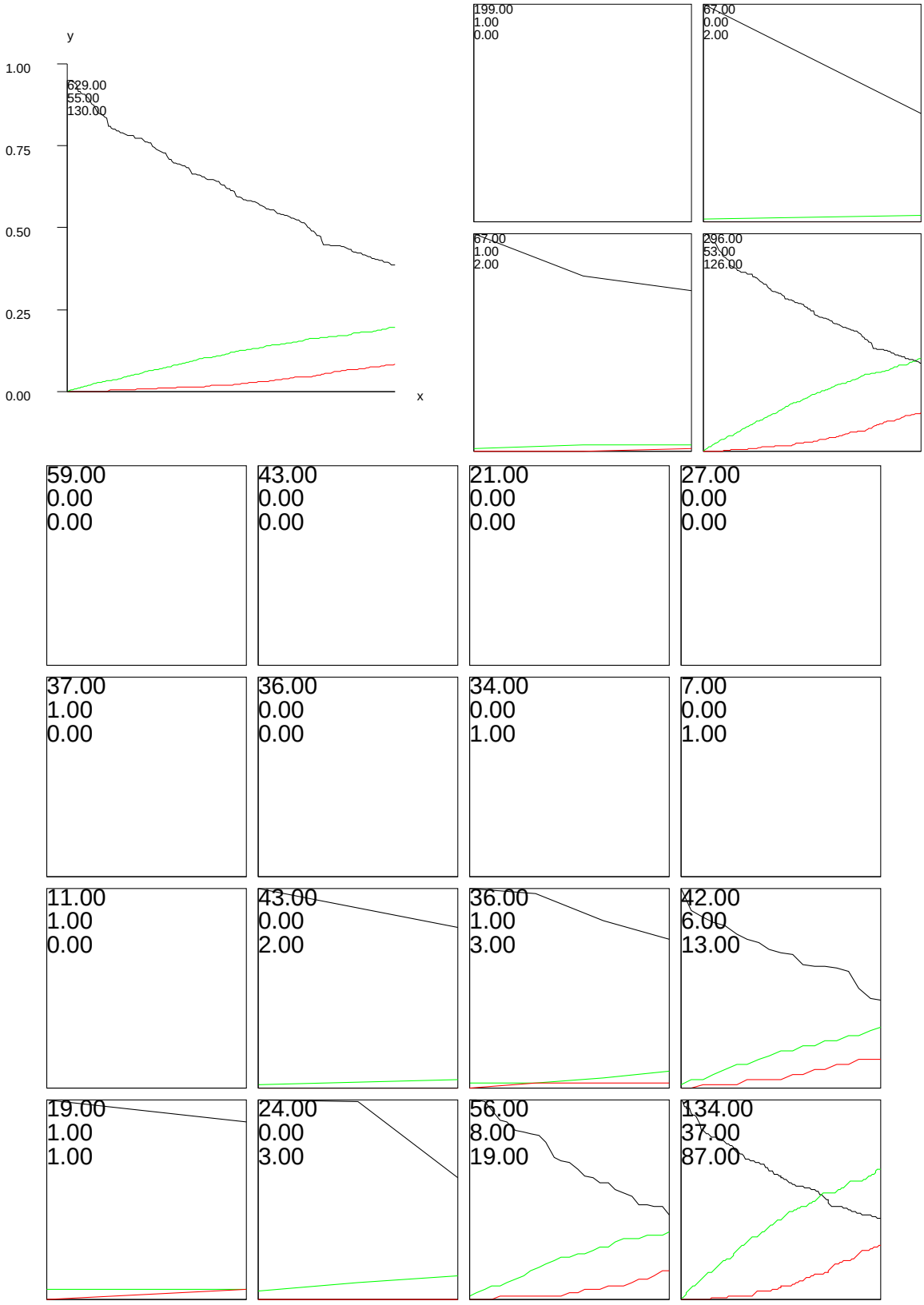


Fig. 6: Error detection graphs for PURTOT (expr. 2)

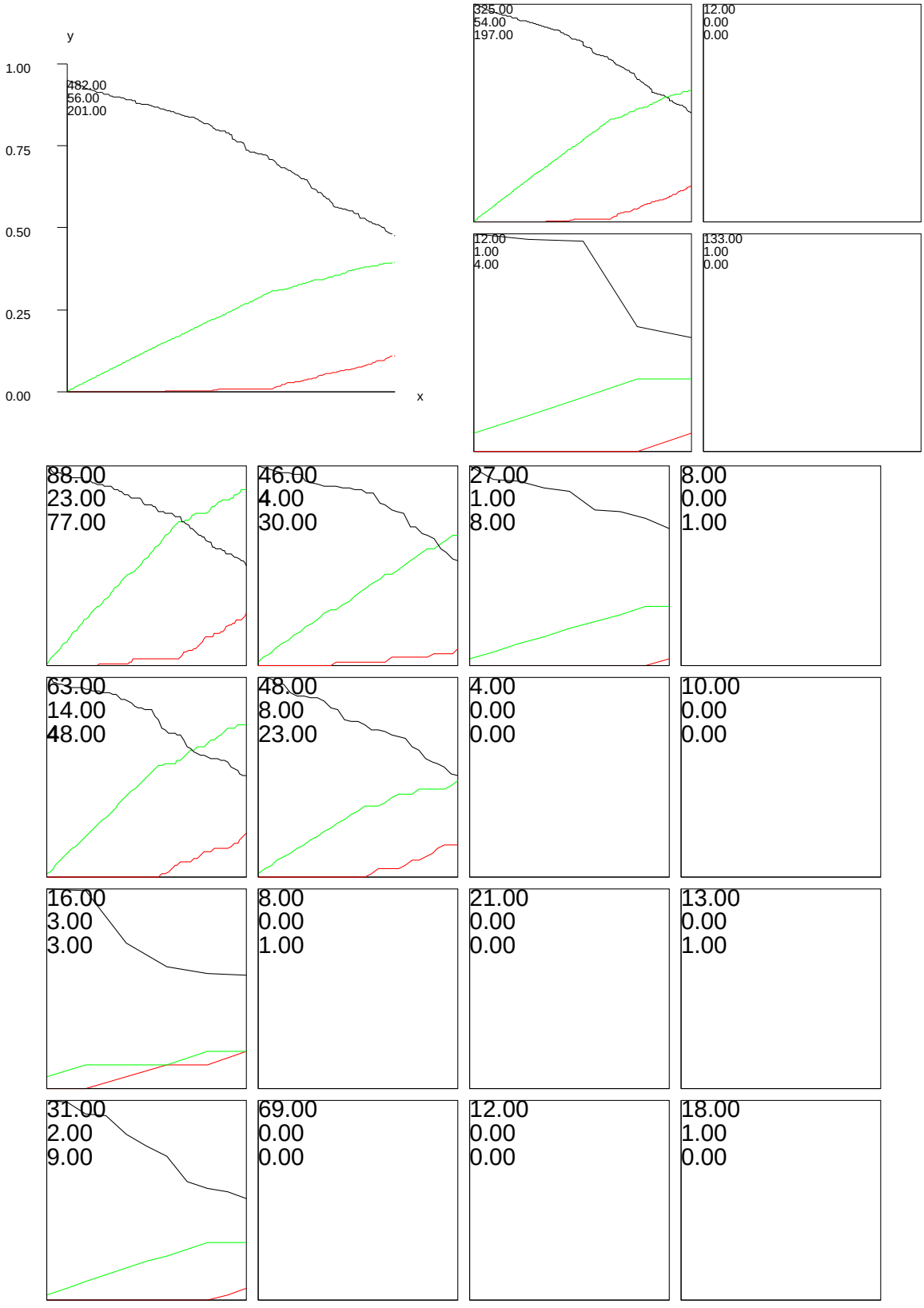


Fig. 7: Error detection graphs for TAXTOT (expr. 1)

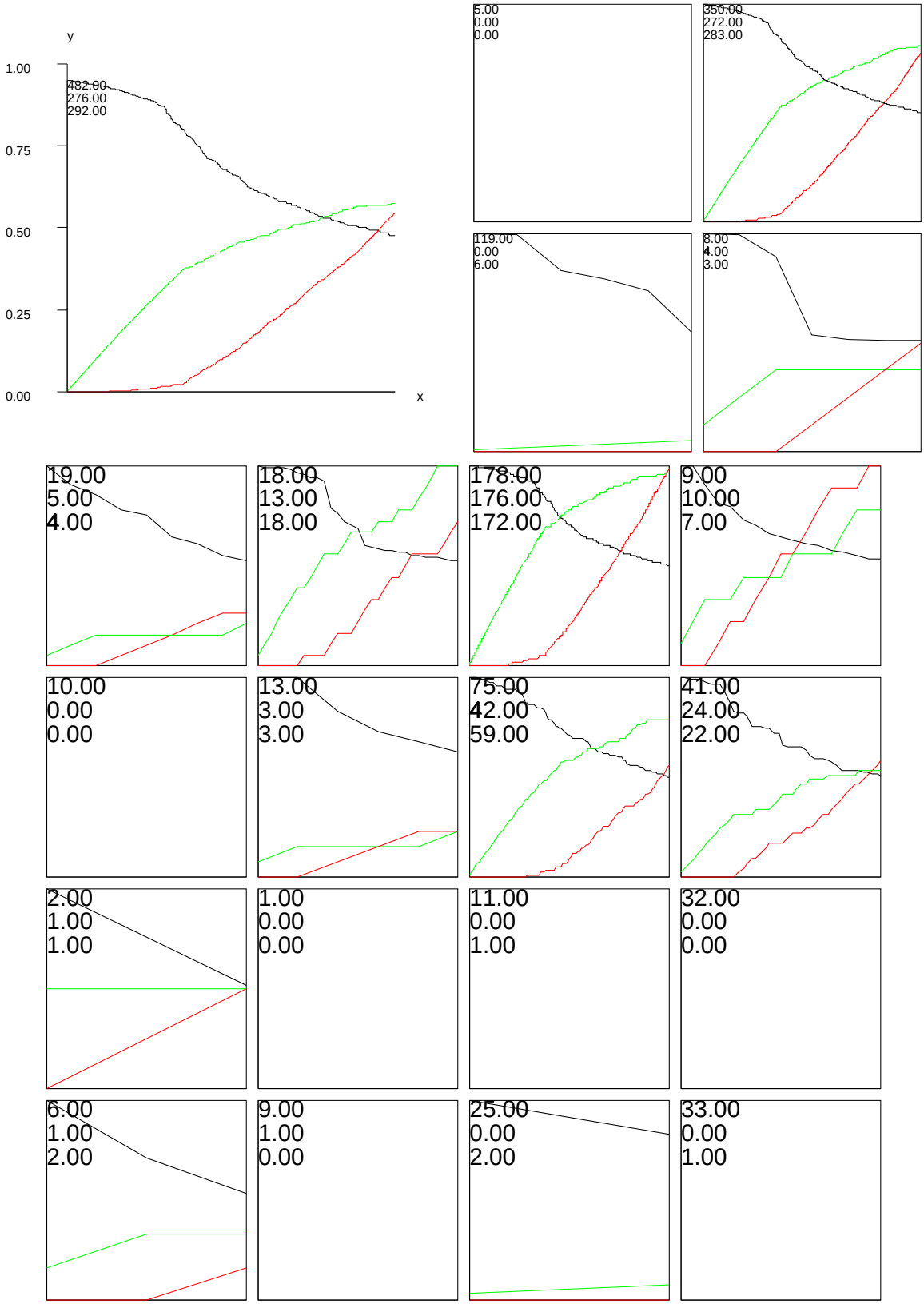


Fig. 8: Error detection graphs for TAXTOT (expr. 2)

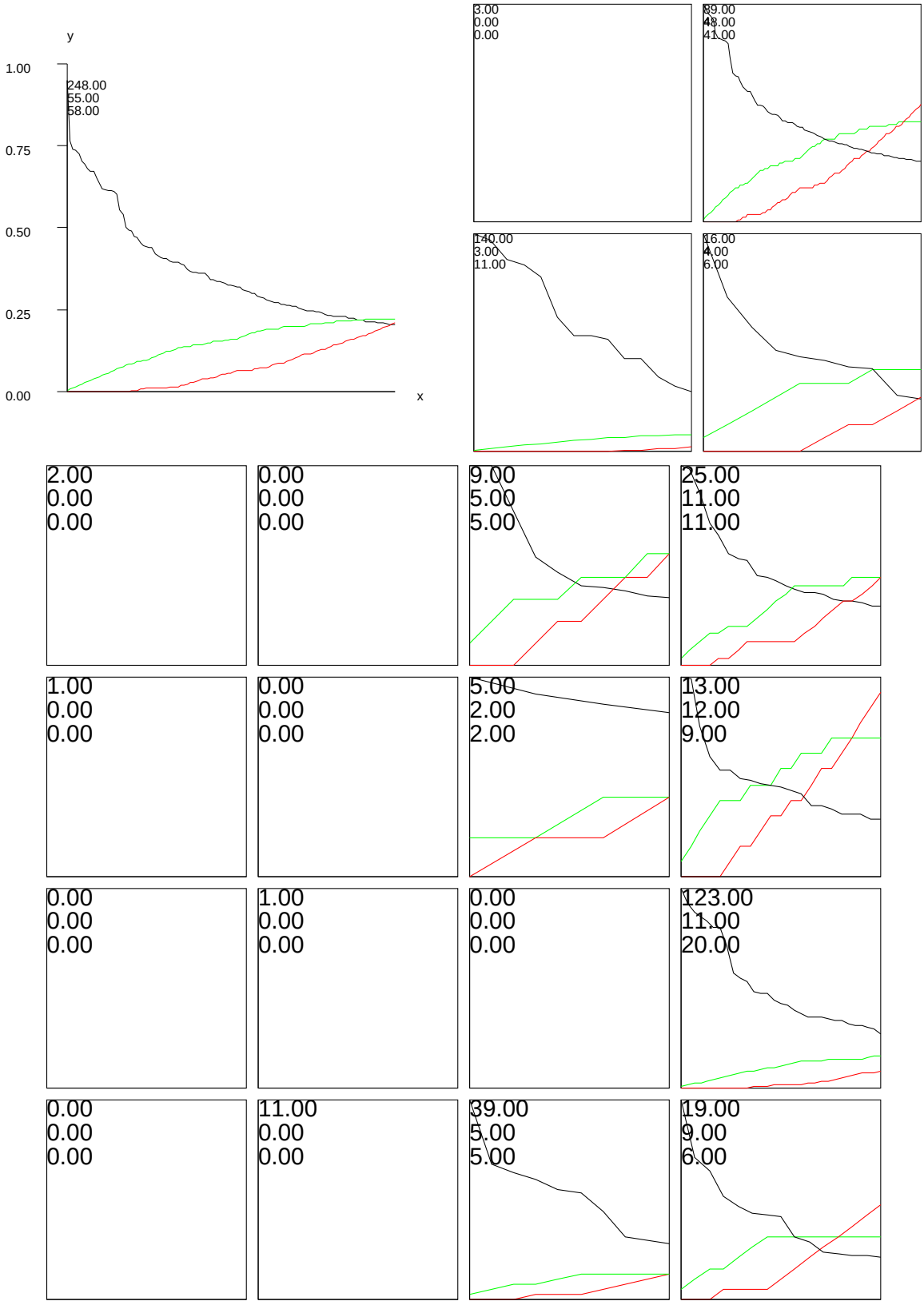


Fig. 9: Error detection graphs for ASSACQ (expr. 1)

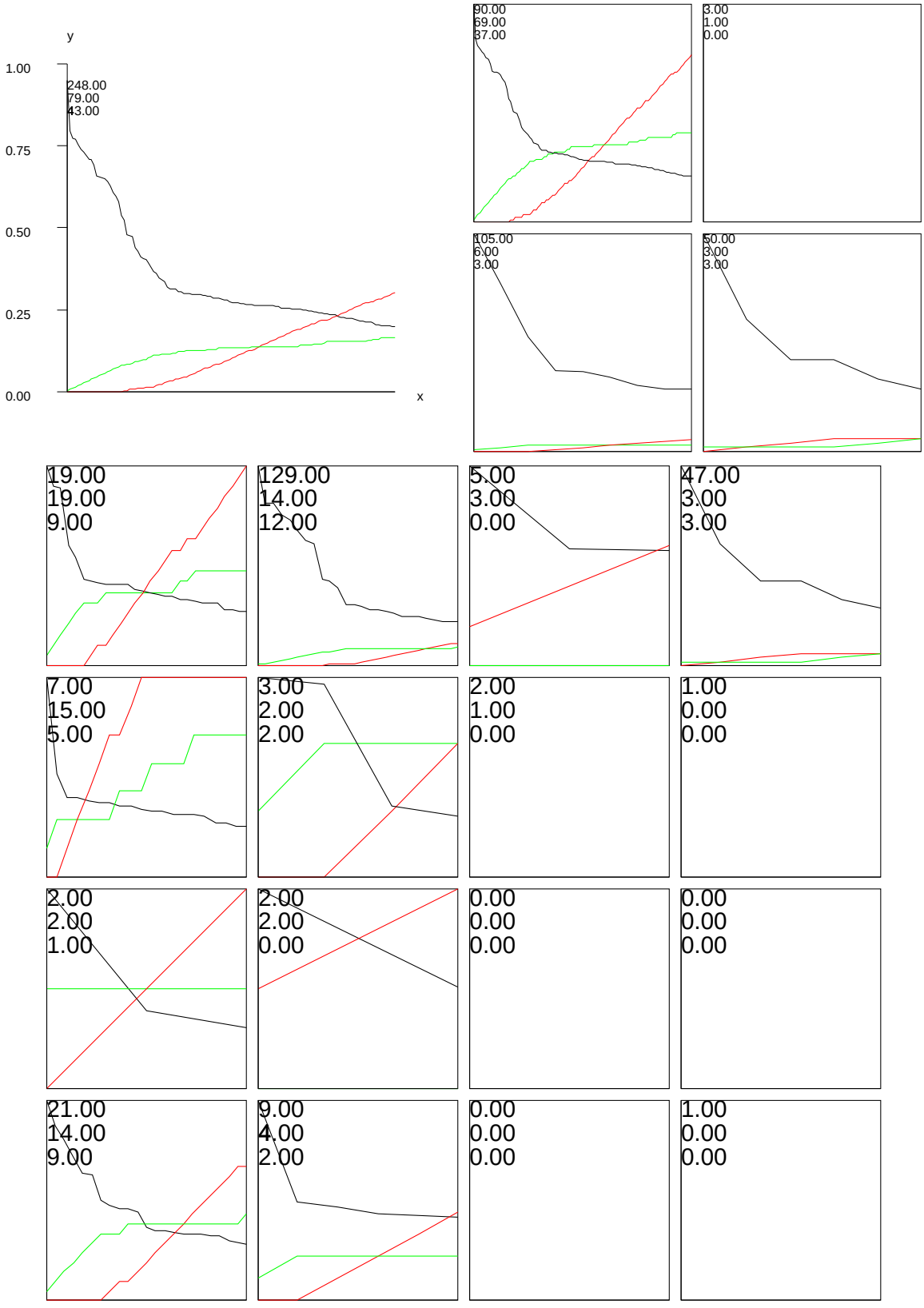


Fig. 10: Error detection graphs for ASSACQ (expr. 2)

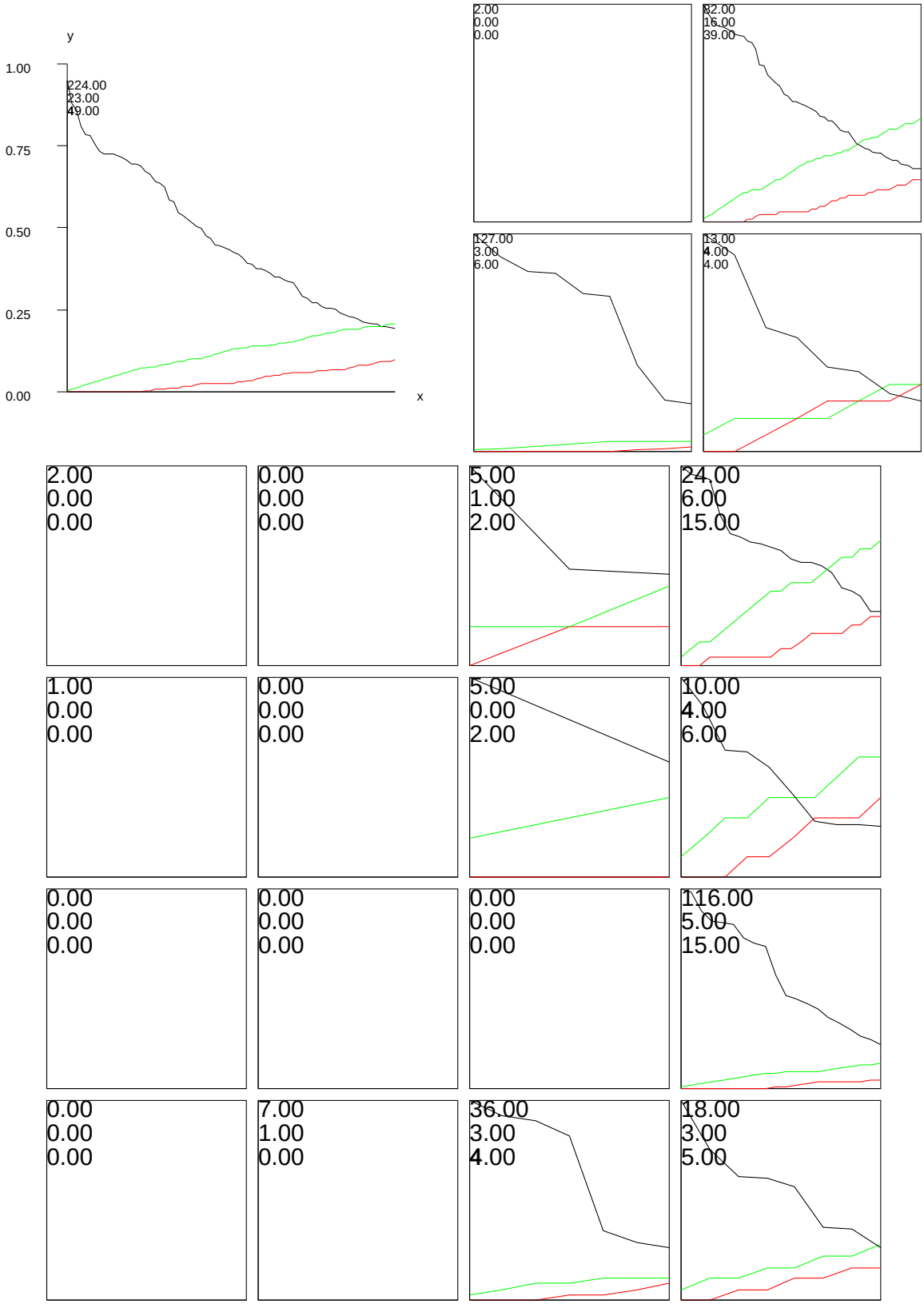


Fig. 11: Error detection graphs for ASSDISP (expr. 1)

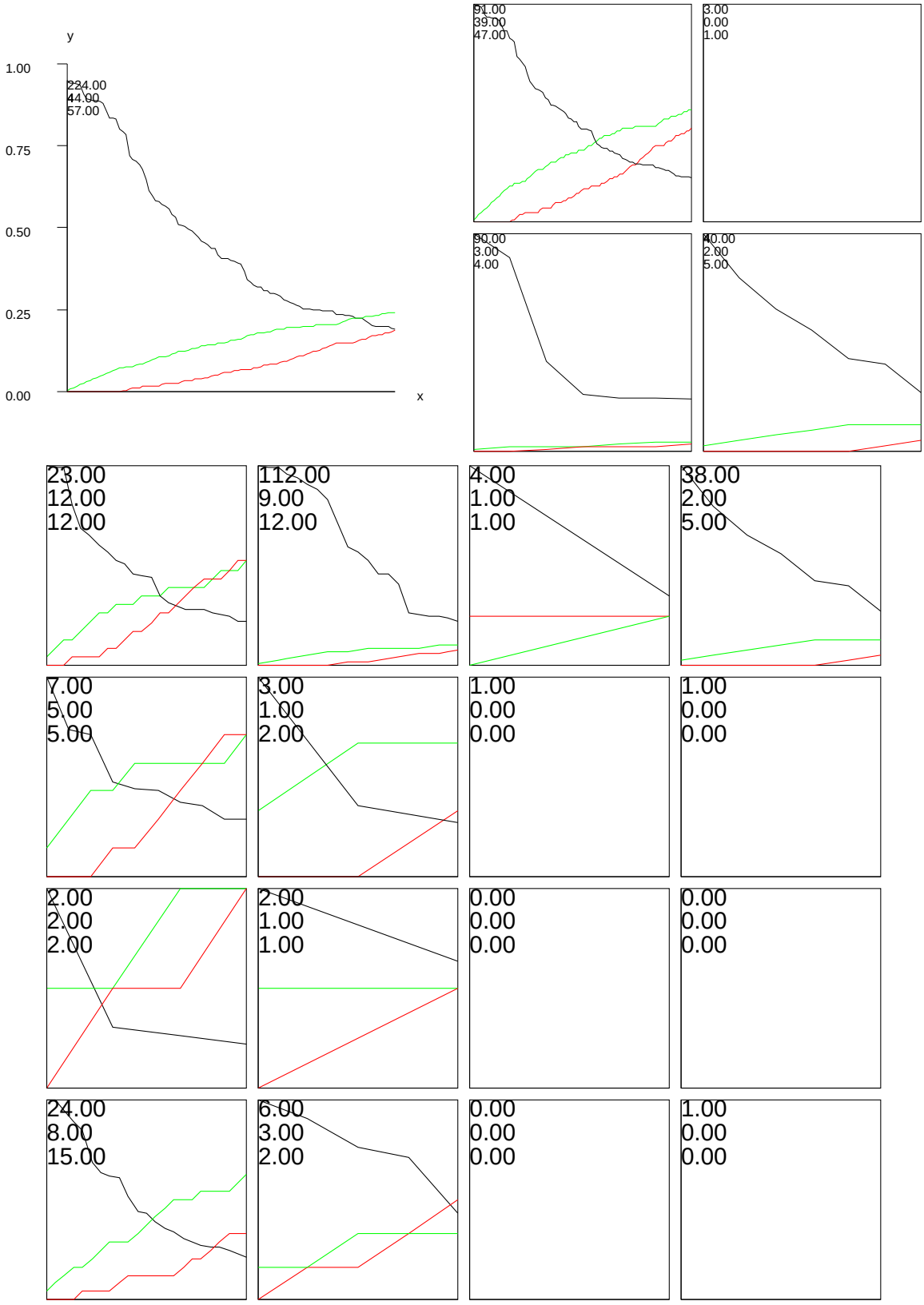


Fig. 12: Error detection graphs for ASSDISP (expr. 2)

5 Configuring TS-SOM edit using UK SARS development datasets

We give an example of detecting editing parameters for two categorial variables which are RELAT (individual) and INSIDEWC (household). Development UK SARS sector Y3 and true datasets were used to detect the editing parameters. The parameters for a categorial variable, or actually for each of its categories, are train and edit cut probabilities. The edit parameters for variables and experiments are shown in table 4.

The chosen two variables have no invalid categories; which are trivial to detect. The explanatory variables for RELAT are SEX, ECONPRIM, ROOMSNUM, PERSINHH, MSTATUS, AGE and QUALEVEL. The INSIDEWC is modelled with variables HHSPTYPE, CENHEAT, BATH, ROOMSNUM, PERSINHH, QUALEVEL, ISCO1, LTILL and HOURS.

We present two experiments as with UK ABI dataset. First one being conservative and second more aggressive when considering error detection. The latter one is derived from the first one by modifying train and edit cut probabilities.

Table 5 shows the error detection statistics. Figures 13-16 have 4 graphs which correspond TS-SOM layers 0-3 (1, 4, 16 and 64 neurons/clusters). Contents of the figures are same as corresponding figures of UK ABI (the earlier section).

Variable	**TrainCutPr/EditCutPr** (expr. 1)	**TrainCutPr/EditCutPr** (expr. 2)
RELAT	0.45/0.5	0.2/0.1
INSIDEWC	0.5/0.5	0.2/0.1

Table 4: Edit train and cut probabilities

Variable	**Expr.**	**N**	**N(TRUE)**	**N(FALSE)**
RELAT	1	2827	2266	1897
	2	2827	2451	2765
INSIDEWC	1	2009	?	?
	2	2009	2007	101

Table 5: SARS Y3 error detection statistics

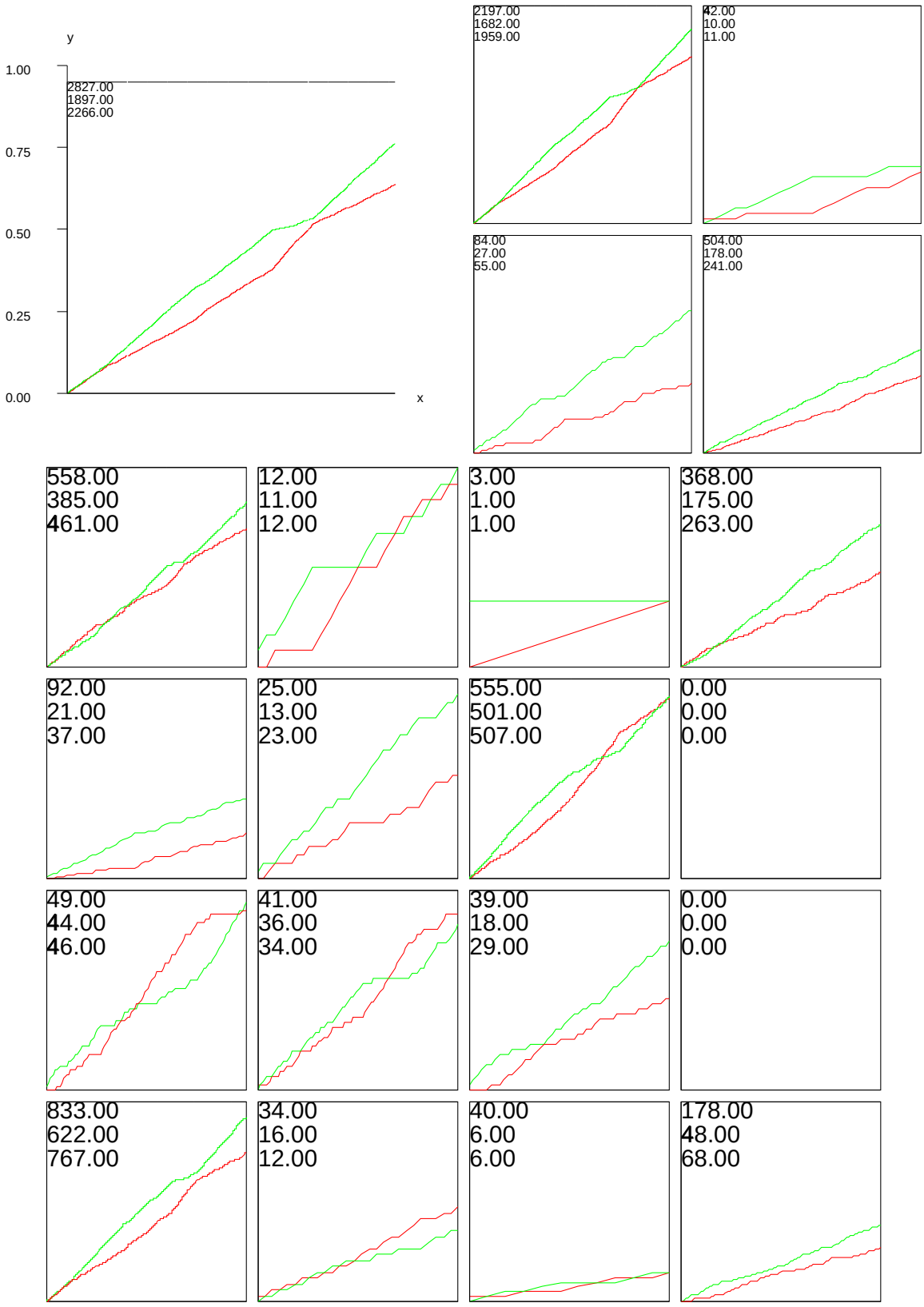


Fig. 13: Error detection graphs 1/2 for RELAT (expr. 1)

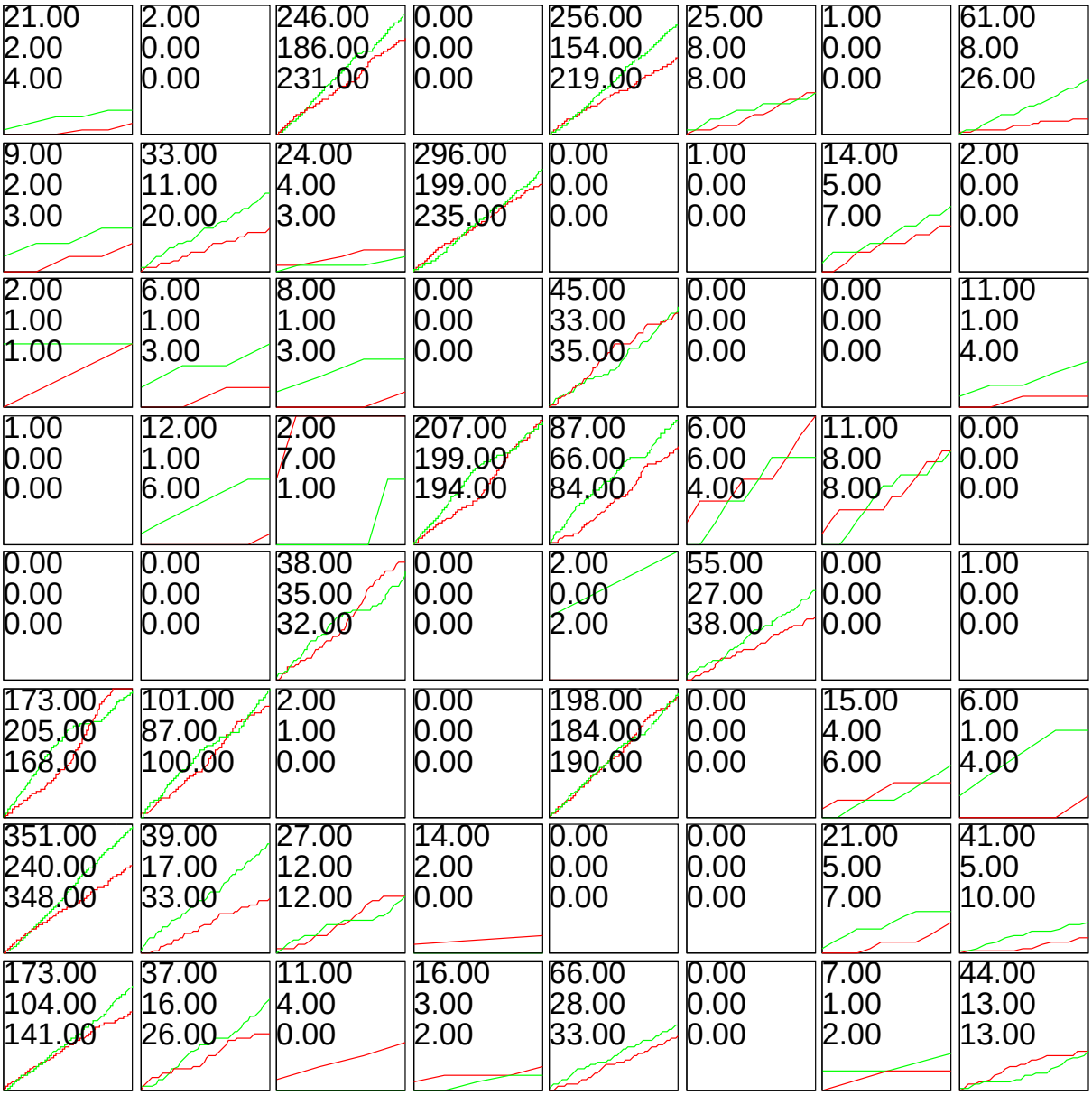


Fig. 13.1: Error detection graphs 2/2 for RELAT (expr. 1)

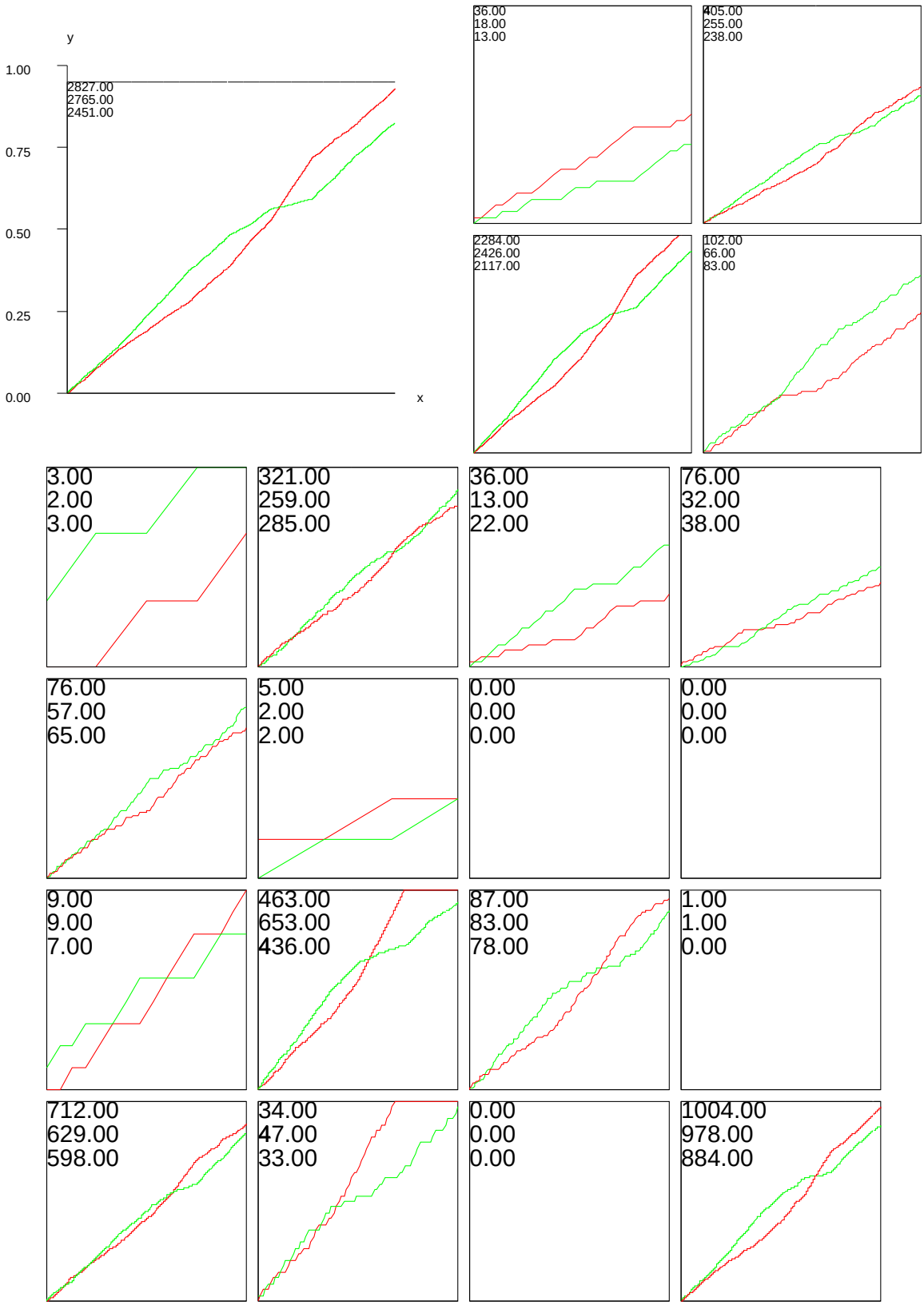


Fig. 14: Error detection graphs 1/2 for RELAT (expr. 2)

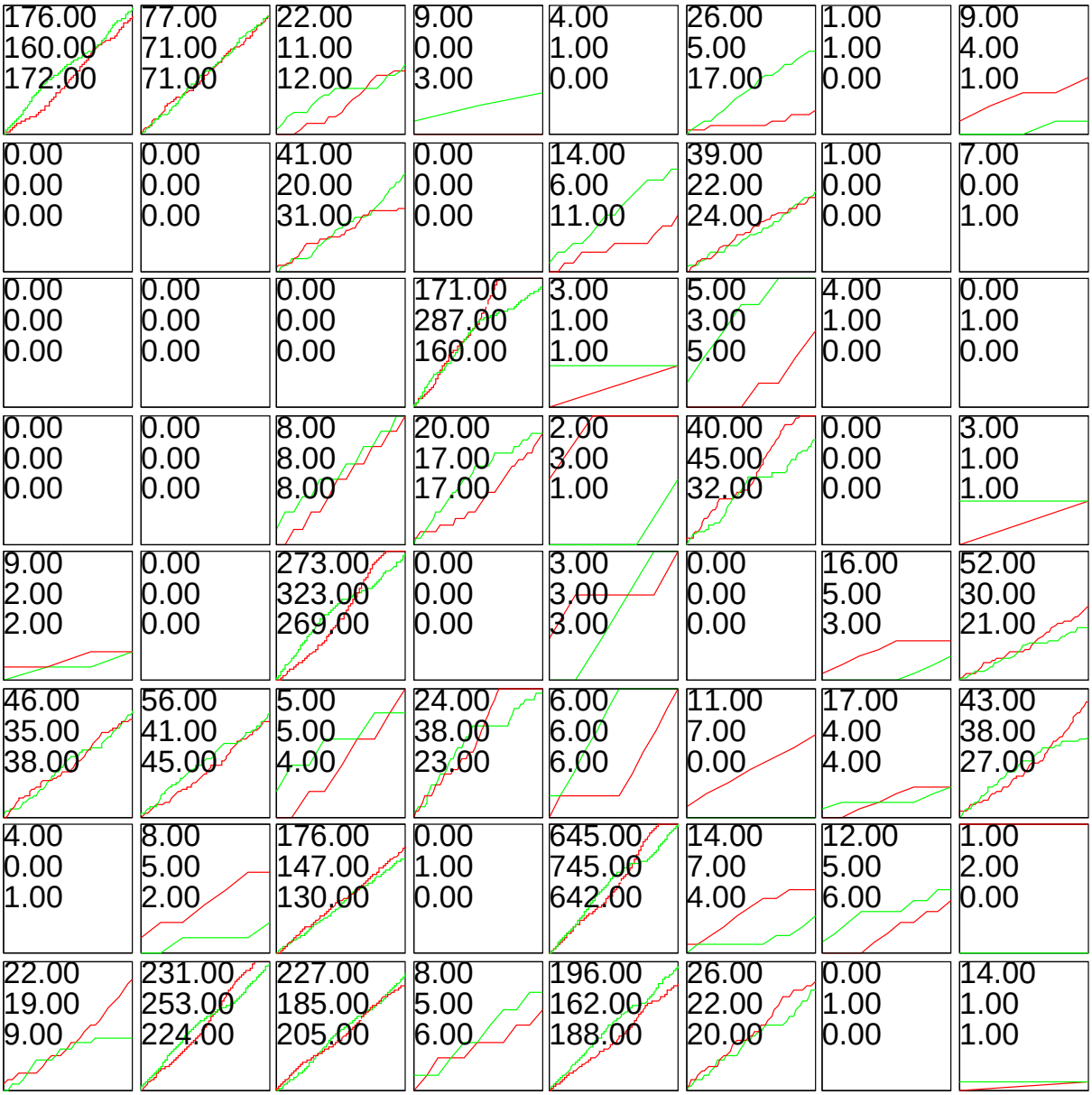


Fig. 14.1: Error detection graphs 2/2 for RELAT (expr. 2)

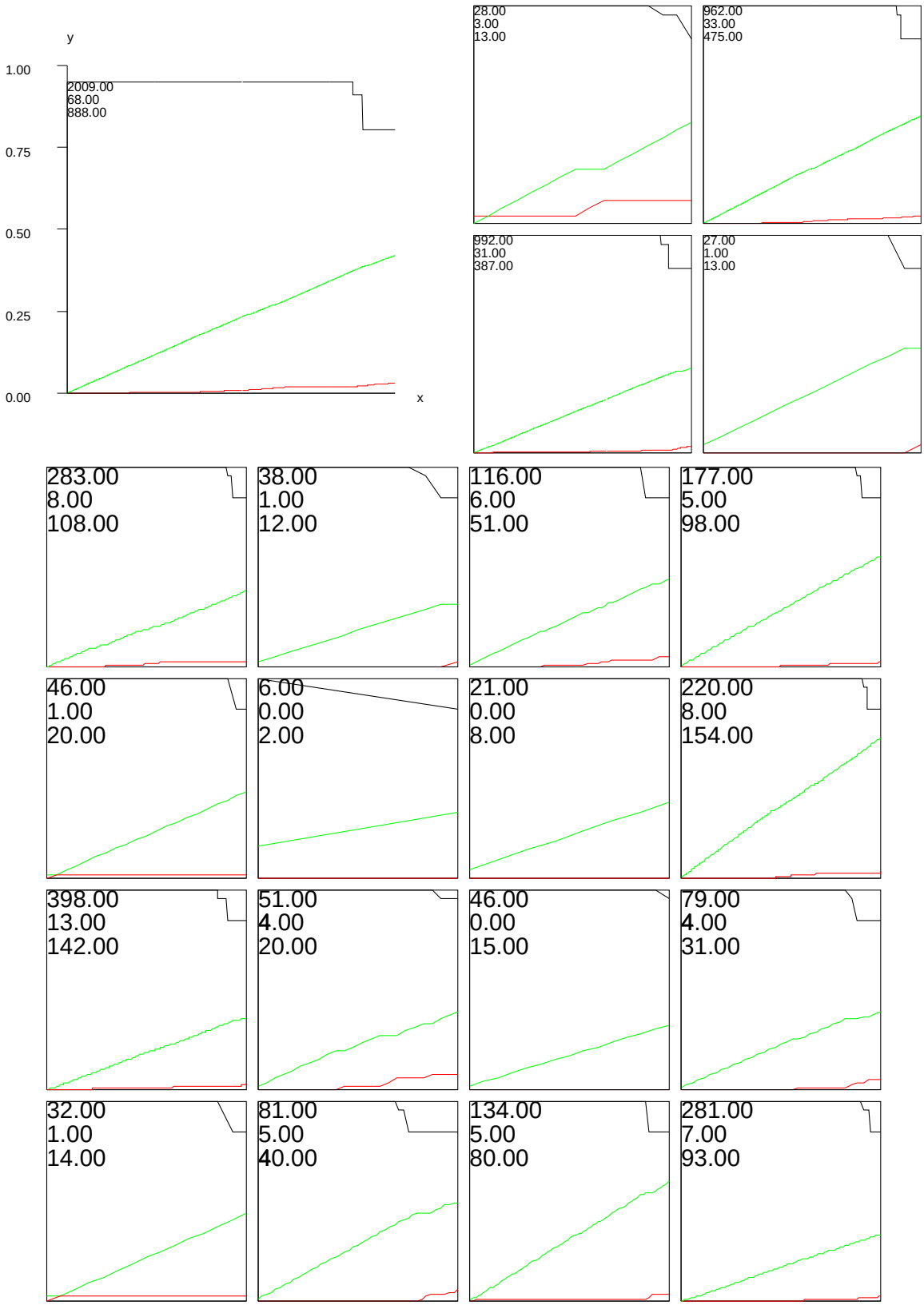


Fig. 15: Error detection graphs 1/2 for INSIDEWC (expr. 1)

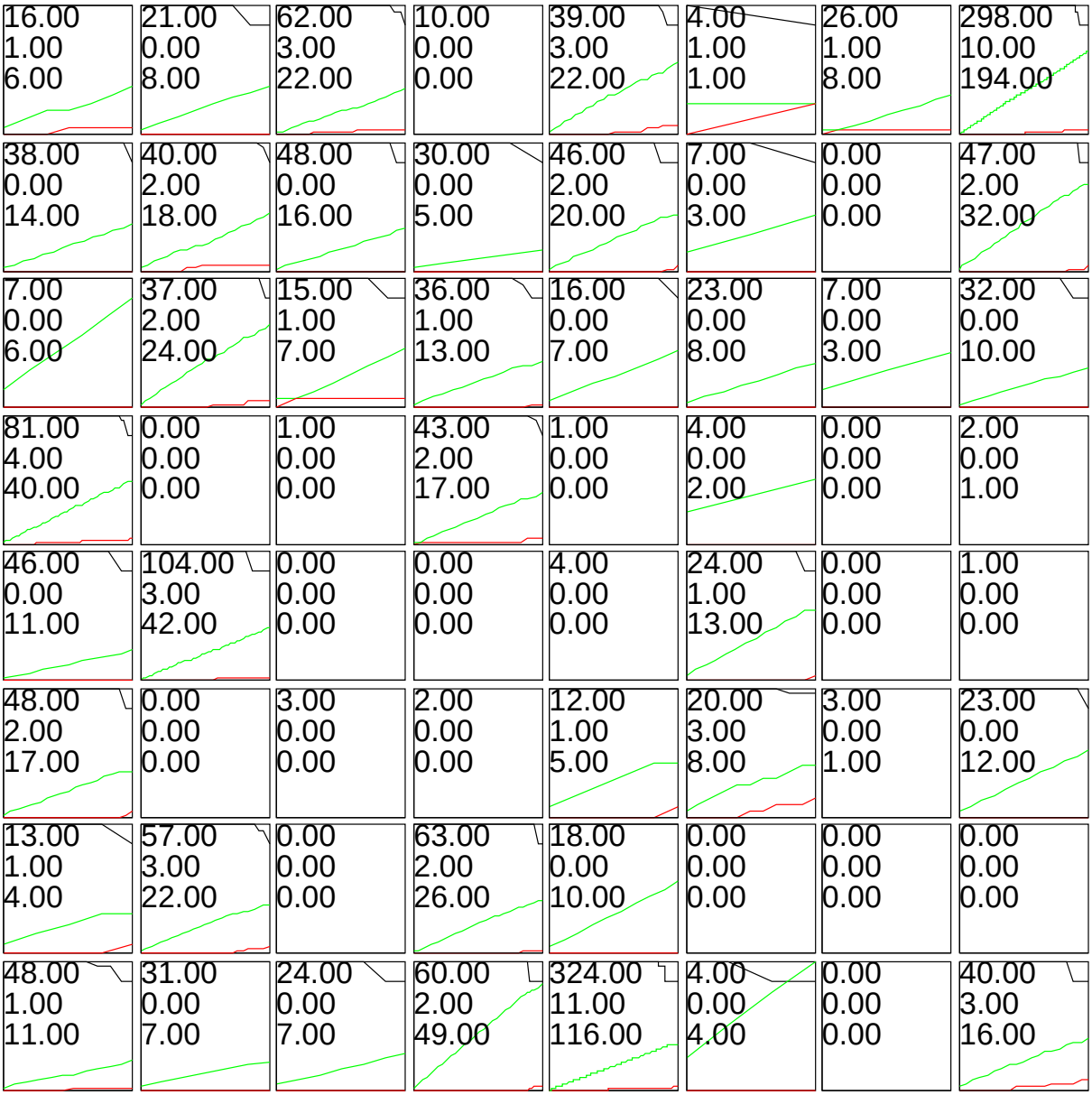


Fig. 15.1: Error detection graphs 2/2 for INSIDEWC (expr. 1)

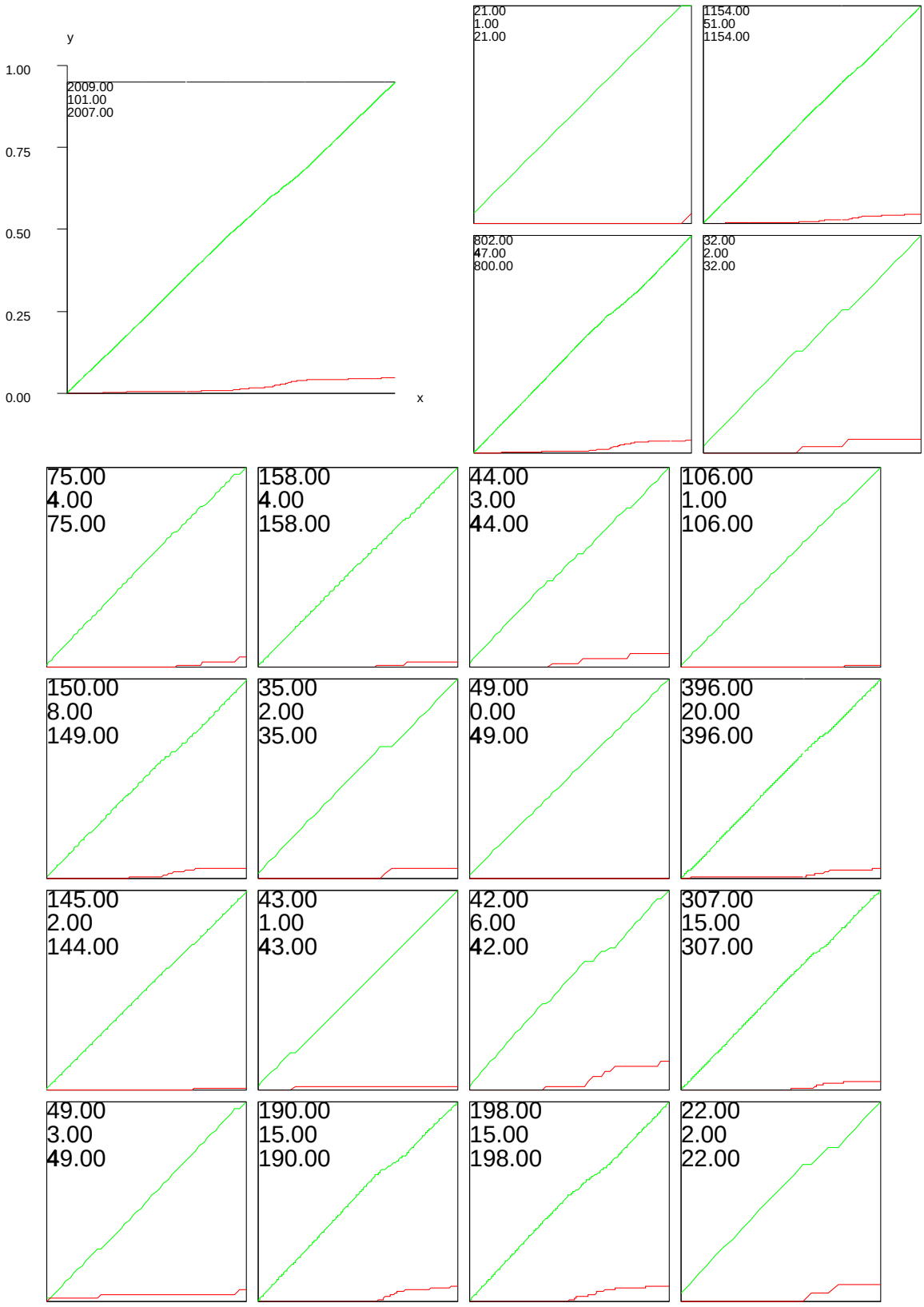


Fig. 16: Error detection graphs 1/2 for INSIDEWC (expr. 2)

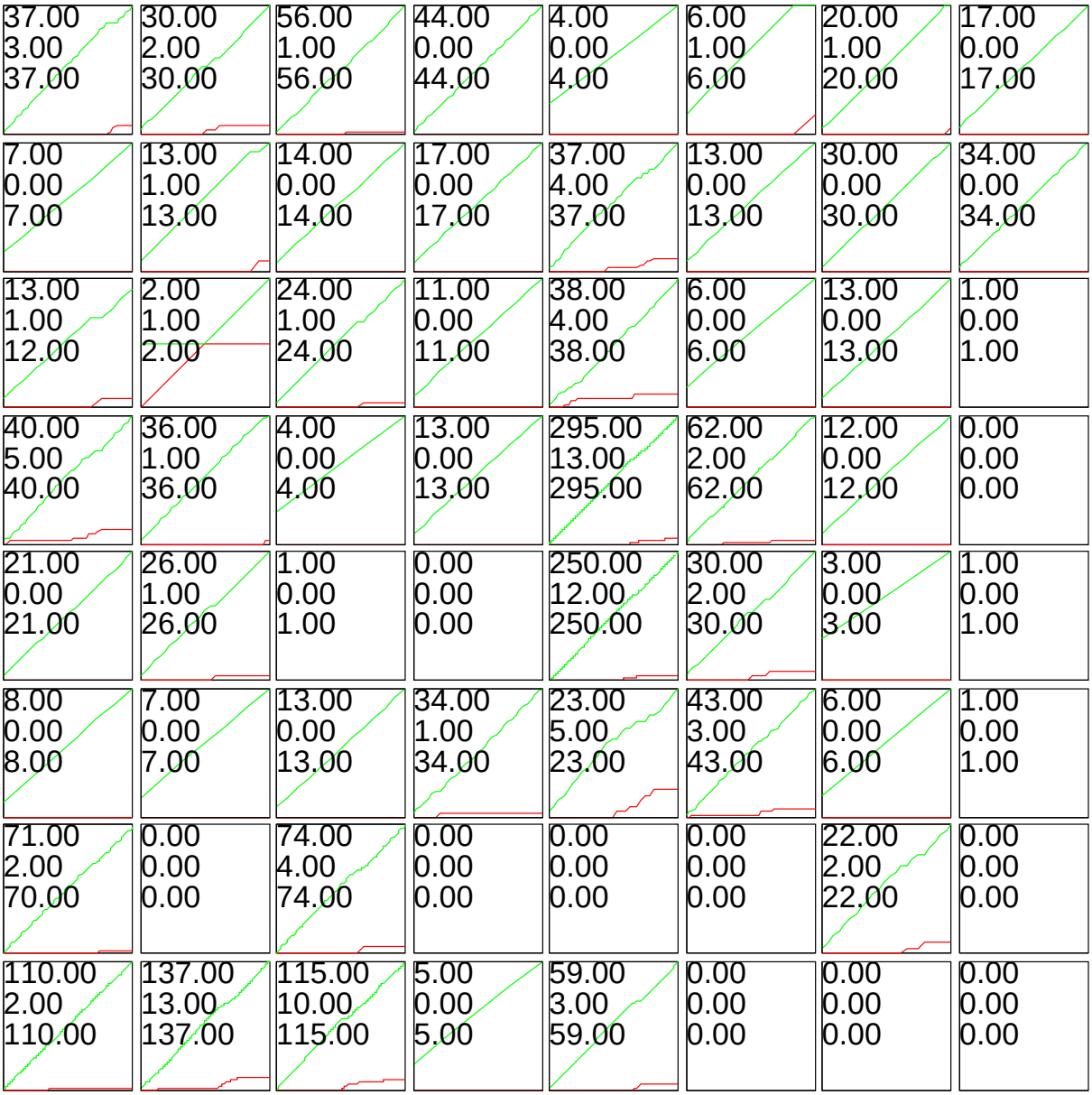


Fig. 16.1: Error detection graphs 2/2 for INSIDEWC (expr. 2)

6 Experiments with all of the data sets: Y2 datasets

In our experiments we imputed all Y2 (incomplete) datasets which were The Danish Labour Force Survey, European Community Household Survey, UK Annual Business Inquiry, UK Sample of Anonymised Records and Swiss Environment Protection Expenditures. We included at least two experiments for each of the datasets. We were mainly interested from the results of pure TS-SOM based imputation because it is our speciality. On the other hand it was noticed that use of external knowledge (logical edit rules) in imputation improves results. So we did some hybrid experiments in which both external knowledge and TS-SOM were used to get as good results as possible.

6.1 The Danish Labour Force Survey (LFS)

A random sample of 15579 records was taken from the development data because the development data was much bigger than the evaluation data. The sample was used to set good TS-SOM imputation layer. Continuous variables were min-max equalized (to range [0,1]) and all categorial variables were binarized.

The incomplete variable INCOME was imputed using explanatory variables AGE, SEX, EDUCATIO, BUSINESS, UNEMPLOY, MARRIAGE and PHONE. The variables were chosen because their combination produced best imputation results with the development dataset.

Experiment	**Options**
1.	A0B0C1D0E0F0G0H0I1J0K0L5
2.	A0B0C1D0E0F0G0H0I1J0K5L2
3.	A0B0C1D0E0F0G0H0I1J0K3L5
4.	A0B0C1D0E0F0G0H0I1J0K1L5
5.	A0B0C1D0E0F0G0H0I1J0K4L3
6.	A0B0C1D0E0F0G0H0I0J0K1L5

Table 6: Options for LFS experiments

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1.	0.92	0.31	0.084072	0.014533	0.000441	1689	6353651065
2.	0.91	0.39	0.081198	0.017584	0.000669	4020	4478491030
3.	0.85	0.20	0.035928	0.008017	0.000115	402	3611230060
4.	0.87	0.23	0.041677	0.00706	0.000088	1982	4377038288
5.	0.86	0.26	0.043353	0.007561	0.000136	947	2533995253
6.	0.92	0.29	0.182754	0.025507	0.001794	1559	7036968412

Table 7: LFS imputation statistics 1/2 for INCOME

Experiment	DL1	DL2	DLINF	MSE
1.	53451	89549	955189	1195908
2.	48972	83607	804359	2150981
3.	64992	104274	955189	1055018
4.	62562	100347	955189	1321713
5.	60267	98870	955189	1121543
6.	54844	89806	846528	1123896

Table 8: LFS imputation statistics 2/2 for INCOME

6.2 European Community Household Survey (GSOEP)

Because of panel structure of the data one of our experiment (MLP) took advantage of the structure. The income of next year was predicted using income of earlier year. However the income of first year was imputed using TS-SOM imputation. In our other experiments the years were modelled and imputed separately.

TS-SOM training variables for INCOME91 and HHINCO91 in experiment 1 are SEX, Y.O.B. (year of birth), BETR91 and ERWTYP91. INCOME92-96 and HHCINCO92-96 are imputed year by year using MLP regression (within TS-SOM clusters). The MLP training variables are Y.O.B, SEX, ERWTYP, INCOME, HHINCO. Latter two variables are from the previous year. The TS-SOM variables for INCOME92-96 and HHINCO92-96 are SEX, Y.O.B, BETR, ERWTYP (from the year being imputed). ...

Experiment/variables	**Options**
Experiment 1	
INCOME91, HHINCO91	A0B0C1D0E0F0G0H0I1J0K0L5
INCOME92-96/HHINCO92-96	A0B0C1D0E0F0G0H0I1J0K5L3
Experiment 2	
INCOME91-96/HHINCO91-96	A0B0C1D0E0F0G0H0I1J0K0L4
Experiment 3	
INCOME91-96/HHINCO91-96	A0B0C1D0E0F0G0H0I1J0K1L4
Experiment 4	
INCOME91-96/HHINCO91-96	A0B0C1D0E0F0G0H0I0J0K1L3

Table 9: Options for GSOEP experiments

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.91	0.57	0.246884	0.015662	0.000903	663	217355973
2	0.93	0.53	0.262908	0.021791	0.001611	598	289820626
3	0.67	0.25	0.091395	0.011983	0.000412	1400	165445914
4	0.63	0.20	0.110979	0.016582	0.000839	2232	187629321

Table 10: GSOEP imputation statistics 1/2 for INCOME91

Experiment	DL1	DL2	DLINF	MSE
1.	11212	18049	213084	252836
2.	11599	18721	213621	240621
3.	18474	28221	192133	421711
4.	20627	29675	216597	719518

Table 11: GSOEP imputation statistics 2/2 for INCOME91

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.97	0.25	0.183976	0.061415	0.007157	279	1293000936
2	0.97	0.25	0.27181	0.070125	0.010071	190	1343949860
3	0.73	0.08	0.059347	0.018279	0.000598	1831	161382661
4	0.60	0.00	0.103264	0.036114	0.002511	1994	743088294

Table 12: GSOEP imputation statistics 1/2 for HHINCO91

...

Experiment	DL1	DL2	DLINF	MSE
1.	26236	36366	216696	422810
2.	26219	36355	221492	414350
3.	37848	49919	235772	838039
4.	48302	61303	261900	946655

Table 13: GSOEP imputation statistics 2/2 for HHINCO91

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.98	0.77	0.329493	0.010079	0.000406	139	177911995
2	0.93	0.57	0.307028	0.022191	0.001567	241	346167016
3	0.73	0.31	0.15841	0.013583	0.000596	1037	19075090
4	0.64	0.22	0.123848	0.018278	0.000863	1541	154627719

Table 14: GSOEP imputation statistics 1/2 for INCOME92

Experiment	DL1	DL2	DLINF	MSE
1.	7675	14025	145050	239096
2.	12023	19150	191842	225532
3.	18094	27322	171374	348486
4.	21593	31023	178147	493856

Table 15: GSOEP imputation statistics 2/2 for INCOME92

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	1.00	0.59	0.14977	0.014262	0.000784	1219	1108778320
2	0.98	0.24	0.243088	0.03252	0.004461	1449	1827672564
3	0.75	0.06	0.060484	0.007047	0.000205	798	543245634
4	0.60	0.01	0.108295	0.01547	0.001	3269	629160549

Table 16: GSOEP imputation statistics 1/2 for HHINCO92

Experiment	DL1	DL2	DLINF	MSE
1.	18496	30093	416532	639466
2.	28668	41065	526967	642304
3.	40969	54519	522080	587843
4.	50718	65725	595999	1690751

Table 17: GSOEP imputation statistics 2/2 for HHINCO92

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.97	0.75	0.310006	0.003387	0.000093	123	245392997
2	0.89	0.46	0.24465	0.011438	0.00078	1010	404966749
3	0.64	0.21	0.100058	0.008911	0.000361	2262	212845859
4	0.61	0.14	0.080393	0.008795	0.000333	1529	77030985

Table 18: GSOEP imputation statistics 1/2 for INCOME93

...

Experiment	DL1	DL2	DLINF	MSE
1.	7387	15736	320238	283341
2.	13223	23231	434860	360582
3.	21114	33724	435278	819606
4.	24127	35136	432046	529013

Table 19: GSOEP imputation statistics 2/2 for INCOME93

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	1.00	0.56	0.153268	0.007898	0.00024	1910	1493017480
2	0.95	0.21	0.249277	0.02679	0.00357	1146	2329598209
3	0.73	0.08	0.049161	0.005818	0.000146	133	731353355
4	0.55	0.01	0.13823	0.017479	0.001571	5243	945655927

Table 20: GSOEP imputation statistics 1/2 for HHINCO93

Experiment	DL1	DL2	DLINF	MSE
1.	17613	35446	618457	968606
2.	30530	47360	710143	645988
3.	42166	59665	679837	628772
4.	56603	74152	677126	3551732

Table 21: GSOEP imputation statistics 2/2 for HHINCO93

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.97	0.82	0.352375	0.006755	0.000295	722	16698435
2	0.90	0.55	0.352375	0.020689	0.001334	909	272628931
3	0.67	0.31	0.130487	0.015156	0.000545	2215	330395834
4	0.57	0.22	0.122069	0.026336	0.001762	5257	671793238

Table 22: GSOEP imputation statistics 1/2 for INCOME94

Experiment	DL1	DL2	DLINF	MSE
1.	7054	13057	136791	381331
2.	12875	20600	183042	388215
3.	19996	30525	159290	799013
4.	24674	35026	210332	2983787

Table 23: GSOEP imputation statistics 2/2 for INCOME94

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	-	-	0.1816	-	-	720	642873649
2	0.95	0.26	0.212267	0.037059	0.00452	721	1917033646
3	0.69	0.08	0.066146	0.010247	0.000347	1253	47903035
4	0.52	0.01	0.165364	0.029451	0.003062	11527	2424567877

Table 24: GSOEP imputation statistics 1/2 for HHINCO94

...

Experiment	DL1	DL2	DLINF	MSE
1.	15965	28838	425751	733486
2.	30013	43657	441125	627911
3.	43418	60212	425089	855927
4.	57710	74118	432056	13475786

Table 25: GSOEP imputation statistics 2/2 for HHINCO94

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.99	0.54	0.347463	0.001844	0.000089	666	1016159212
2	0.92	0.32	0.347463	0.005474	0.000399	44	1285398653
3	0.71	0.20	0.176119	0.002904	0.000111	1105	757364973
4	0.62	0.09	0.133731	0.003822	0.000204	2464	619673158

Table 26: GSOEP imputation statistics 1/2 for INCOME95

Experiment	DL1	DL2	DLINF	MSE
1.	8007	29842	1078871	334385
2.	14845	36120	1141982	265977
3.	21236	41588	1125074	406499
4.	26194	46492	1162948	880254

Table 27: GSOEP imputation statistics 2/2 for INCOME95

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	1.02	0.50	0.199403	0.004821	0.000145	2950	2556622606
2	0.96	0.16	0.215522	0.016084	0.001889	2489	3605092000
3	0.73	0.06	0.078806	0.00414	0.000119	2434	2179799336
4	0.58	0.01	0.094925	0.007917	0.000457	3960	559535068

Table 28: GSOEP imputation statistics 1/2 for HHINCO95

Experiment	DL1	DL2	DLINF	MSE
1.	16508	45402	1117261	1459227
2.	32624	58695	1167911	1123959
3.	43998	69466	1155134	1202996
4.	55329	80951	1210388	2216284

Table 29: GSOEP imputation statistics 2/2 for HHINCO95

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	-	-	0.334895	-	-	198	354428597
2	0.91	0.45	0.369438	0.012928	0.001049	150	719376566
3	0.70	0.25	0.162178	0.006476	0.000278	669	244066053
4	0.54	0.11	0.106557	0.009172	0.000477	2521	126032739

Table 30: GSOEP imputation statistics 1/2 for INCOME96

...

Experiment	DL1	DL2	DLINF	MSE
1.	8349	17841	370372	319399
2.	15317	27385	547600	289072
3.	22138	34908	532102	350415
4.	28541	42438	553804	966096

Table 31: GSOEP imputation statistics 2/2 for INCOME96

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	-	-	0.226581	-	-	924	864695591
2	0.95	0.26	0.26171	0.036238	0.005394	328	2143465730
3	0.68	0.08	0.119438	0.008578	0.000317	500	490469010
4	0.53	0.02	0.11007	0.015884	0.00114	7000	1310176505

Table 32: GSOEP imputation statistics 1/2 for HHINCO96

Experiment	DL1	DL2	DLINF	MSE
1.	16216	28823	326944	787811
2.	33293	46743	554593	589448
3.	45707	62339	535935	731646
4.	57324	74085	491505	5714071

Table 33: GSOEP imputation statistics 2/2 for HHINCO96

...

6.3 UK Annual Business Inquiry (ABI)

With some of the experiments the external knowledge (logical edit rules) were used to derive formulas for deterministic imputation. Approximate 20% of the missing data values were imputed that way. Short form non applicable (-9) values were handled as incomplete data. Continuous variables were log transformed and min-max equalized. Categorial variables were binarized.

Experiment/variables	**Options**
Experiment 1	
TURNOVER,EMPTOTC,PURTOT	A0B1C1D0E0F0G0H0I1J0K0L4
TAXTOT	A0B1C1D0E0F0G0H0I1J0K0L2
ASSACQ,ASSDISP	A0B1C1D0E0F0G0H0I1J0K4L2
Experiment 2	
TURNOVER,EMPTOTC,PURTOT	A0B1C1D0E2F0G0H0I1J0K0L4
TAXTOT	A0B1C1D0E2F0G0H0I1J0K0L2
ASSACQ,ASSDISP	A0B1C1D0E2F0G0H0I1J0K4L2
Experiment 3	
TAXTOT,EMPTOTC,TURNOVER,PURTOT	A0B1C1D0E0F0G0H0I1J0K5L2
ASSACQ,ASSDISP	A0B1C1D0E0F0G0H0I1J0K4L2
Experiment 4	
TAXTOT,EMPTOTC,TURNOVER,PURTOT	A0B1C1D0E2F0G0H0I1J0K5L2
ASSACQ,ASSDISP	A0B1C1D0E2F0G0H0I1J0K4L2
Experiment 5	
TURNOVER,EMPTOTC,PURTOT,	A0B1C1D0E0F0G0H0I0J0K1L4
TAXTOT,ASSACQ,ASSDISP	
Experiment 6	
TURNOVER,EMPTOTC,PURTOT,	A0B1C1D0E2F0G0H0I0J0K1L4
TAXTOT,ASSACQ,ASSDISP	

Table 34: Options for ABI experiments

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	1.09	0.35	0.066176	0.002108	0.000033	703	1025962653
2	0.71	0.05	0.073529	0.001355	0.000024	592	983216153
3	0.99	0.05	0.080882	0.001105	0.000016	612	986266683
4	1.86	0.97	0.102941	0.00337	0.000031	371	739974774
5	0.41	0.09	0.051471	0.002143	0.000027	469	912065878
6	0.72	0.96	0.073529	0.005185	0.000056	7	860512271

Table 35: ABI imputation statistics 1/2 for TURNOVER

Experiment	DL1	DL2	DLINF	MSE
1	813	30833	50522	1724815
2	895	31295	50779	1670953
3	700	31276	52037	1677869
4	595	15596	24154	1612186
5	1037	30605	48071	1726226
6	557	13425	20083	1750159

Table 36: ABI imputation statistics 2/2 for TURNOVER

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.97	0.63	0.092437	0.002572	0.00006	10.8	102234
2	1.00	0.95	0.117647	0.002117	0.000037	5.3	3881
3	0.63	0.49	0.134454	0.004438	0.000161	3.0	54091
4	1.00	0.95	0.12605	0.002079	0.000099	4.8	3134
5	0.74	0.53	0.092437	0.003101	0.000089	7.9	97213
6	0.95	0.96	0.07563	0.009081	0.00011	0.6	1769

Table 37: ABI imputation statistics 1/2 for EMPTOTC

Experiment	DL1	DL2	DLINF	MSE
1	28.8	253.6	370.9	12601
2	18.8	81.8	67.8	9905
3	37.8	247.7	299.8	13054
4	27.1	77.0	20.5	11712
5	35.4	257.6	350.5	13904
6	25.6	70.6	33.6	10537

Table 38: ABI imputation statistics 2/2 for EMPTOTC

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.66	0.17	0.128571	0.00669	0.000089	101.2	51092476
2	-	-	0.035714	0.006582	0.000047	104.1	47785643
3	1.30	0.56	0.364286	0.004455	0.000062	217.5	67182605
4	-	-	0.057143	0.006348	0.000046	99.9	47473720
5	0.73	0.14	0.078571	0.005756	0.000065	148.4	68634339
6	1.00	0.58	0.021429	0.007969	0.000105	91.3	46320172

Table 39: ABI imputation statistics 1/2 for PURTOT

Experiment	DL1	DL2	DLINF	MSE
1	296.4	7658.3	8997.4	2067121
2	108.6	6382.0	9974.7	301751
3	286.2	7229.4	9612.5	2416054
4	107.0	6148.8	9610.2	301752
5	302.4	8014.4	10650.7	2415973
6	105.6	5603.2	8756.9	344305

Table 40: ABI imputation statistics 2/2 for PURTOT

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.01	0.63	0.102362	0.00579	0.000077	13.8	11405
2	0.01	0.29	0.181102	0.004382	0.000193	38.2	17339
3	1.14	0.91	0.094488	0.00626	0.000091	1.9	2670
4	1.00	0.98	0.07874	0.001512	0.000038	1.3	352
5	0.46	0.24	0.204724	0.003774	0.000061	1.1	6339
6	0.44	0.18	0.19685	0.008044	0.000083	0.7	6194

Table 41: ABI imputation statistics 1/2 for TAXTOT

Experiment	DL1	DL2	DLINF	MSE
1	19.8	87.7	47.5	129
2	44.1	134.7	17.6	154
3	4.4	29.3	36.0	152
4	3.0	12.2	9.6	127
5	8.6	79.2	114.8	209
6	8.5	82.7	125.2	203

Table 42: ABI imputation statistics 2/2 for TAXTOT

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	6.70	1.00	0.21	0.001678	0.000016	130.4	65944173
2	6.70	0.98	0.215	0.001517	0.000015	128.1	65803099
3	6.70	0.96	0.195	0.001762	0.000018	132.9	65980998
4	6.70	0.96	0.21	0.00173	0.000017	134.1	65996088
5	0.73	0.01	0.11	0.001132	0.000007	139.4	67195280
6	0.00	0.00	0.175	0.001257	0.00001	141.1	67269897

Table 43: ABI imputation statistics 1/2 for ASSACQ

Experiment	DL1	DL2	DLINF	MSE
1	140.0	6990.2	13481.1	3310
2	135.0	6949.6	13481.1	3131
3	141.4	7032.0	13481.1	2875
4	143.2	7036.9	13481.1	3251
5	161.4	8178.3	15816.5	1919
6	173.7	8208.1	15834.6	1956

Table 44: ABI imputation statistics 2/2 for ASSACQ

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.03	0.04	0.355263	0.001628	0.000112	0.3	24687
2	0.20	0.41	0.388158	0.001094	0.000125	1.6	22419
3	0.14	0.01	0.342105	0.001865	0.000108	0.9	23162
4	0.06	0.00	0.375	0.001633	0.000121	0.5	24677
5	-	-	0.197368	0.004026	0.000037	1.6	11330
6	-	-	0.171053	0.004535	0.000039	0.9	16921

Table 45: ABI imputation statistics 1/2 for ASSDISP

Experiment	DL1	DL2	DLINF	MSE
1	7.7	155.7	265.3	134
2	7.7	131.6	220.4	146
3	7.8	158.7	268.2	125
4	7.8	157.5	268.5	132
5	6.0	169.9	219.6	90
6	9.6	255.7	332.9	90

Table 46: ABI imputation statistics 2/2 for ASSDISP

...

6.4 UK Sample of Anonymised Records (SARS)

The data contained many categorial variables and only a few continuous variables. All categorial variables were binarized and continuous variables were min-max equalized.

Experiment/variables	**Options**
Experiment 1 AGE,ROOMSNUM SEX,MSTATUS,LTILL,RELAT, HHSPTYPE,BATH,CENHEAT, INSIDEWC,TENURE, CARS	 A0B0C1D0E0F0G0H0I0J0K0L5 A0B0C0D0E0F0G0H0I0J0K0L5
Experiment 2 AGE,ROOMSNUM SEX,MSTATUS,LTILL,RELAT, HHSPTYPE,BATH,CENHEAT, INSIDEWC,TENURE, CARS	 A0B0C1D0E0F0G0H0I0J0K3L5 A0B0C0D0E0F0G0H0I0J0K3L5
Experiment 3 AGE,ROOMSNUM SEX,MSTATUS,LTILL,RELAT, HHSPTYPE,BATH,CENHEAT, INSIDEWC,TENURE, CARS	 A0B0C1D0E0F0G0H0I0J0K1L7 A0B0C0D0E0F0G0H0I0J0K0L7

Table 47: Options for SARS experiments

...

Experiment	W	D	Eps
1	2.612903	0.525424	0.346051
2	0.043478	0.389831	0.186441
3	3.333333	0.508475	0.325927

Table 48: SARS imputation statistics for SEX

...

Experiment	W	D	Eps
1	25.213152	0.589286	0.418006
2	4.630759	0.428571	0.226541
3	24.525526	0.589286	0.418006

Table 49: SARS imputation statistics for RELAT

...

Experiment	W	D	Eps
1	12.908397	0.361111	0.172713
2	4.02439	0.222222	0.014352
3	10.082645	0.319444	0.125

Table 50: SARS imputation statistics for MSTATUS

...

Experiment	W	D	Eps
1	0	0.129032	0
2	4	0.064516	0
3	0.666667	0.096774	0

Table 51: SARS imputation statistics for LTILL

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.38	0.04	0.451613	0.234946	0.075087	21.6	1816
2	0.74	0.11	0.258065	0.071825	0.010898	4.3	133
3	0.39	0.01	0.580645	0.263082	0.106197	24.2	1823

Table 52: SARS imputation statistics 1/2 for AGE

Experiment	DL1	DL2	DLINF	MSE
1	26.7	33.6	73	0.001063
2	10.1	14.1	46	0.001056
3	26.8	33.0	68	0.001065

Table 53: SARS imputation statistics 2/2 for AGE

...

Experiment	W	D	Eps
1	14.298873	0.702703	0.523426
2	6.334448	0.75	0.573223
3	18.073826	0.815789	0.676539

Table 54: SARS imputation statistics for TENURE

...

Experiment	W	D	Eps
1	4.270981	0.586207	0.417277
2	1.383035	0.586207	0.417277
3	3.953349	0.551724	0.375896

Table 55: SARS imputation statistics for HHSPTYPE

...

Experiment	W	D	Eps
1	21.409654	0.6	0.448814
2	4.907937	0.857143	0.766792
3	24.603724	0.628571	0.482885

Table 56: SARS imputation statistics for ROOMSNUM

...

Experiment	W	D	Eps
1	1	0.017857	0
2	0.150019	0.107143	0
3	0	0.035714	0

Table 57: SARS imputation statistics for BATH

...

Experiment	W	D	Eps
1	14.86901	0.414634	0.245654
2	6.267806	0.487805	0.329738
3	15.307692	0.45122	0.287605

Table 58: SARS imputation statistics for CENHEAT

...

Experiment	W	D	Eps
1	2	0.05	0
2	1	0.025	0
3	0	0	0

Table 59: SARS imputation statistics for INSIDEWC

...

Experiment	W	D	Eps
1	0.966655	0.415385	0.22571
2	1.598792	0.415385	0.22571
3	1.09929	0.430769	0.243607

Table 60: SARS imputation statistics for CARS

...

6.5 Swiss Environment Protection Expenditures (EPE)

The data contains huge amount of zero values. Therefore the data was partitioned to 8 subgroups, according to variables netinv, curexp, receipts and subsid. The subgroups correspond to: some net investments, no net investments, some expenditures, no expenditures, some receipts, no receipts, some subsidies and no subsidies. Investments, expenditures, receipts and subsidiers variables were imputed in corresponding subgroups. The partitioning was done because of high amount of zeroes in the data.

Continuous variables were log transformed and min-max equalized and categorial variables were binarized. In second experiment of ours edit rules were used to derive formulas for deterministic imputation. The deterministic imputation was used to impute approximate 20% of missing data values.

Experiment/variables	**Options**
Experiment 1	
TOTINVTO,TOTEXPTO,RECTOT	A0B1C1D0E0F0G0H0I0J0K1L3
SUBTOT	A0B1C1D0E0F0G0H0I0J0K1L2
Experiment 2	
TOTINVTO,TOTEXPTO,RECTOT	A0B1C1D0E2F0G0H0I0J0K1L3
SUBTOT	A0B1C1D0E2F0G0H0I0J0K1L2

Table 61: Options for EPE experiments

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.17	0.05	0.166667	0.009083	0.000287	54.5	177812
2	0.60	0.06	0.222222	0.021334	0.001495	16.7	2939

Table 62: EPE evaluation statistics 1/2 for TOTINVTO

Experiment	DL1	DL2	DLINF	MSE
1	169.5	534.7	676.7	-
2	100.5	376.5	468.6	411

Table 63: EPE evaluation statistics 2/2 for TOTINVTO

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.02	0.01	0.091429	0.011272	0.000256	164.6	742598
2	-	-	0.04	0.001081	0.00002	3.0	18470

Table 64: EPE evaluation statistics 1/2 for TOTEXPTO

Experiment	DL1	DL2	DLINF	MSE
1	241.9	934.8	605.8	-
2	7.1	44.1	87.5	52

Table 65: EPE evaluation statistics 2/2 for TOTEXPTO

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.03	0.01	0.214286	0.024275	0.001476	56.3	15009
2	0.03	0.00	0.261905	0.01241	0.000623	34.9	14955

Table 66: EPE evaluation statistics 1/2 for RECTOT

Experiment	DL1	DL2	DLINF	MSE
1	85.2	135.1	81.4	1.4
2	65.1	137.0	99.0	1.1

Table 67: EPE evaluation statistics 2/2 for RECTOT

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	-	-	0.5	0	0	7.7	169
2	-	-	0.5	0	0	7.7	169

Table 68: EPE evaluation statistics 1/2 for SUBTOT

Experiment	DL1	DL2	DLINF	MSE
1	7.7	11.1	10.2	0.08
2	7.7	11.1	10.2	0.08

Table 69: EPE evaluation statistics 2/2 for SUBTOT

7 Y3 datasets

There were three Y3 (incomplete and erroneous) datasets which were UK Annual Business Inquiry (ABI), UK Sample of Anonymised Records (SARS) and Swiss Environment Protection Expenditures (EPE). We did not process the last one because we run out of time.

7.1 UK Annual Business Inquiry (ABI)

All continuous variables were log transformed and robustly min-max equalized (using 5% and 95% fractiles). Categorial variables were binarized. Short form non applicable (-9) values were handled as incomplete data.

There was information that amount of errors in the evaluation data is higher than in the development data. We did three experiments of which first one was conservative, second more aggressive and last the most aggressive when considering error detection percentage. First experiment included outlier imputation, in two other experiments outliers were just detected and marked as incomplete data.

Experiment/variables	**Options**
Experiment 1 TURNOVER,EMPTOTC,PURTOT, TAXTOT,ASSACQ,ASSDISP	A1B1C1D0E0F0G1H0I0J2K1L2
Experiment 2 TURNOVER,EMPTOTC,PURTOT, TAXTOT,ASSACQ,ASSDISP	A1B1C1D0E0F0G1H0I0J0K1L2
Experiment 3 TURNOVER,EMPTOTC,PURTOT, TAXTOT,ASSACQ,ASSDISP	A1B1C1D0E0F0G1H0I0J0K1L2

Table 70: Options for SARS (y3) experiments

Experiment	G	A	B	C
1	0.040373	0.471827	0.233691	0.300818
2	0.039195	0.36938	0.39723	0.389379
3	0.179945	0	1	0.718113

Table 71: ABI (y3) evaluation global error statistics

Experiment	alpha	beta	delta	RAE	RRASE
1	0.646729	0.000883	0.056613	0.185233	0.040259
2	0.614953	0.022948	0.074032	0.06651	0.018875
3	0.04486	0.965225	0.885806	-0.000072	0.000009

Table 72: ABI (y3) error statistics 1/2 for TURNOVER

Experiment	RER	tj	AREm1	AREm2
1	3446.292135	1.379062	0.07347	0.706196
2	1832.258427	0.495164	0.231765	0.874023
3	0.261798	-0.000533	0.81473	0.999835

Table 73: ABI (y3) error statistics 2/2 for TURNOVER

...

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.99	0.01	0.277533	0.010153	0.000216	8027.426687	64575380000
2	1.02	0.13	0.363636	0.010828	0.001243	84.160798	863315
3	1.86	0.15	0.636364	0.019573	0.00501	3.3	27436

Table 74: ABI (y3) imputation statistics 1/2 for TURNOVER

Experiment	DL1	DL2	DLINF	MSE
1	8206.7	254083.6	378957.1	-
2	131	895.3	198.1	-
3	4.0	160.6	6.3	706765

Table 75: ABI (y3) imputation statistics 2/2 for TURNOVER

Experiment	alpha	beta	delta	RAE	RRASE
1	0.677365	0.000536	0.065298	0.309546	0.090986
2	0.60473	0.018231	0.074349	0.132205	0.035025
3	0.043919	0.966041	0.877808	-0.00025	0.000026

Table 76: ABI (y3) error statistics 1/2 for EMPTOTC

Experiment	RER	tj	AREm1	AREm2
1	12737.73529	2.483676	0.171089	0.366148
2	2524.698529	1.060763	0.145996	0.757686
3	0.507353	-0.002007	0.89158	0.999903

Table 77: ABI (y3) error statistics 2/2 for EMPTOTC

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	8.06	0.02	0.5	0.010113	0.000195	554.6	326348524
2	2.58	0.08	0.521739	0.054118	0.007958	9.1	5665
3	9.90	0.05	0.73913	0.057104	0.010419	0.3	168

Table 78: ABI (y3) imputation statistics 1/2 for EMPTOTC

Experiment	DL1	DL2	DLINF	MSE
1	554.9	18063.9	25273.9	-
2	9.4	74.1	17.4	-
3	0.3	12.9	0.5	8815

Table 79: ABI (y3) imputation statistics 2/2 for EMPTOTC

...

Experiment	alpha	beta	delta	RAE	RRASE
1	0.751627	0.001327	0.112921	0.273478	0.051092
2	0.726681	0.015539	0.12131	0.103703	0.028949
3	0.032538	0.963616	0.825133	-0.000075	0.000008

Table 80: ABI (y3) error statistics 1/2 for PURTOT

...

Experiment	RER	tj	AREm1	AREm2
1	4778.750842	1.971625	0.005672	0.652244
2	3032.941077	0.747639	0.205185	0.842369
3	0.272727	-0.000543	0.84947	0.99989

Table 81: ABI (y3) error statistics 2/2 for PURTOT

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	1.15	0.02	0.222222	0.008721	0.000157	5032.2	29760850000
2	1.31	0.14	0.294118	0.004209	0.000453	27.5	421947
3	1.52	0.04	0.617647	0.006191	0.001223	1.0	10802

Table 82: ABI (y3) imputation statistics 1/2 for PURTOT

Experiment	DL1	DL2	DLINF	MSE
1	5120.6	172479.4	271917.7	-
2	52.6	624.7	275.0	-
3	1.4	102.6	7.1	461016

Table 83: ABI (y3) imputation statistics 2/2 for PURTOT

...

Experiment	alpha	beta	delta	RAE	RRASE
1	0.605691	0.012271	0.08293	0.144071	0.009705
2	0.570461	0.055495	0.116812	0.050992	0.001876
3	0.235772	0.561172	0.522427	-0.001496	0.000105

Table 84: ABI (y3) error statistics 1/2 for TAXTOT

Experiment	RER	tj	AREm1	AREm2
1	480	1.36463	0.390971	0.973298
2	99.083333	0.482991	0.63767	0.99853
3	2.166667	-0.014173	0.861067	0.999723

Table 85: ABI (y3) error statistics 2/2 for TAXTOT

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	6.72	0.05	0.661578	0.024272	0.002227	169.9	7809120
2	3.86	0.02	0.542857	0.007571	0.000731	2.6	3313
3	3.14	0.00	0.685714	0.007853	0.001	0.1	176

Table 86: ABI (y3) imputation statistics 1/2 for TAXTOT

Experiment	DL1	DL2	DLINF	MSE
1	170.0	2794.3	4142.5	-
2	2.7	57.5	25.2	-
3	0.1	13.3	1.3	-

Table 87: ABI (y3) imputation statistics 2/2 for TAXTOT

...

Experiment	alpha	beta	delta	RAE	RRASE
1	0.869379	0.000878	0.066667	2.052212	0.109135
2	0.69379	0.008424	0.060341	0.396044	0.061413
3	0.400428	0.422253	0.4206	-0.014204	0.001005

Table 88: ABI (y3) error statistics 1/2 for ASSACQ

Experiment	RER	tj	AREm1	AREm2
1	8716.333333	10.438904	1.699978	0.397144
2	5572.47619	2.014542	0.186458	0.824579
3	51.238095	-0.072251	1.059919	0.999167

Table 89: ABI (y3) error statistics 2/2 for ASSACQ

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	2.33	0.33	0.38806	0.019664	0.000602	742.36385	291955264
2	0.40	0.00	0.338235	0.016725	0.001348	72.673024	280027
3	0.66	0.02	0.573529	0.016922	0.001491	3.7	14300

Table 90: ABI (y3) imputation statistics 1/2 for ASSACQ

Experiment	DL1	DL2	DLINF	MSE
1	746.0	17083.3	17175.5	-
2	76.5	529.0	91.0	-
3	3.9	119.5	4.6	-

Table 91: ABI (y3) imputation statistics 2/2 for ASSACQ

Experiment	alpha	beta	delta	RAE	RRASE
1	0.705405	0.005851	0.047727	0.990602	0.038126
2	0.656757	0.008949	0.047727	0.257895	0.017153
3	0.394595	0.424195	0.422424	-0.030338	0.002173

Table 92: ABI (y3) error statistics 1/2 for ASSDISP

Experiment	RER	tj	AREm1	AREm2
1	341	1.061597	1.067767	0.995182
2	98.888889	0.276378	1.90118	0.997536
3	7.444444	-0.032513	3.266446	0.999914

Table 93: ABI (y3) error statistics 2/2 for ASSDISP

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.48	0.01	0.512821	0.005509	0.000564	131.5	15828547
2	0	0.02	0.730769	0.018053	0.001625	1.0	16
3	-	-	0.826923	0.023417	0.005286	0.7	1

Table 94: ABI (y3) imputation statistics 1/2 for ASSDISP

7.2 UK Sample of Anonymised Records (SARS)

Categorial variables were binarized and continuous variables were robustly min-max equalized. Out of bounds and invalid category errors were detected at first without TS-SOM and their error probabilities were set to 100%. Rest of the errors were detected using TS-SOM.

Experiment	DL1	DL2	DLINF	MSE
1	136.1	3978.0	7962.8	-
2	1.6	4.5	2.6	-
3	0.1	1.0	0.1	-

Table 95: ABI (y3) imputation statistics 2/2 for ASSDISP

Percentage of errors in the evaluation data was assumed to be same as in development data. There was also some evidence of that; the amount of out of bounds and invalid category errors increases with same ratio as the amount of records. Therefore cutting probabilities and sigma parameters for the evaluation data were set so that error detect percentage matched the development data.

TS-SOM edit (and imputation) layer had to be increased with some of the variables for the evaluation dataset because too high portion of the records were marked as erroneous at first. Record(/error) suspect percentage in first experiment was set to match development data in which the percentage is 2%. Outliers were detected and imputed in the the experiment. The experiment 2 and 3 had increased suspect percentage of 2,5%. They also excluded outlier imputation; outliers were just detected and marked as incomplete data.

Some of the variables (e.g. SEX, QUALNUM, QUALEVEL, LTILL, DISTWORK) in the development data had high amount of out of bound and invalid category errors, which are trivial to detect, of all errors in the variable. Therefore if the same is true for the evaluation data then when considering error detection effectiveness of (TS-SOM) algorithm this fact must be taken into account.

Experiment	alpha	beta	delta	Dcat	tj
1	0.111111	0	0.007494	0.007494	0.011202
2	0.111111	0.001786	0.009159	0.007494	0.011202
3	0.111111	0.001786	0.009159	0.007494	0.011202

Table 96: SARS (y3) error statistics for SEX

Experiment	W	D	Eps
1	62.820513	0.478528	0.365404
2	6.211268	0.430303	0.312784
3	1.97561	0.450549	0.295141

Table 97: SARS (y3) imputation statistics for SEX

Experiment	alpha	beta	delta	RAE	RRASE
1	0.744681	0.010909	0.068677	0.00007	0.000015
2	0.712766	0.036364	0.089615	0.000109	0.000018
3	0.712766	0.036364	0.089615	0.000109	0.000018

Table 98: SARS (y3) error statistics 1/2 for AGE

Experiment	RER	tj	AREm1	AREm2
1	2.861111	0.080628	424.293714	424.3179
2	2.944444	0.125442	435.980897	436.02102
3	2.944444	0.125442	435.980897	436.02102

Table 99: SARS (y3) error statistics 2/2 for AGE

...

Experiment	Slope	R^2	K-S	K-S_1	K-S_2	m_1	m_2
1	0.80	0.59	0.41791	0.068342	0.011571	5	253
2	0.75	0.43	0.406061	0.094061	0.017859	7.4	427
3	0.79	0.55	0.295918	0.078776	0.013998	4.5	188

Table 100: SARS (y3) imputation statistics 1/2 for AGE

Experiment	DL1	DL2	DLINF	MSE
1	6.8	10.9	57	0.001072
2	9.4	13.7	58	0.001099
3	5.6	9.8	36	0.001084

Table 101: SARS (y3) imputation statistics 2/2 for AGE

...

Experiment	alpha	beta	delta	Dcat	tj
1	0.23913	0.042254	0.057003	0.017915	0.010776
2	0.195652	0.042254	0.053746	0.014658	0.008817
3	0.195652	0.042254	0.053746	0.014658	0.008817

Table 102: SARS (y3) error statistics for RELAT

Experiment	W	D	Eps
1	79.77753	0.736264	0.66013
2	77.371692	0.704301	0.624557
3	4.440455	0.578125	0.415745

Table 103: SARS (y3) imputation statistics for RELAT

Experiment	alpha	beta	delta	Dcat	tj
1	0.365854	0	0.012788	0.012788	0.018069
2	0.365854	0	0.012788	0.012788	0.018069
3	0.365854	0	0.012788	0.012788	0.018069

Table 104: SARS (y3) error statistics for MSTATUS

Experiment	W	D	Eps
1	2.812489	0.393103	0.263713
2	2.812489	0.393103	0.263713
3	2.825809	0.386555	0.242958

Table 105: SARS (y3) imputation statistics for MSTATUS

Experiment	alpha	beta	delta	Dcat	tj
1	0.162162	0	0.010067	0.010067	0.009566
2	0.162162	0	0.010067	0.010067	0.009566
3	0.162162	0	0.010067	0.010067	0.009566

Table 106: SARS (y3) error statistics for LTILL

Experiment	W	D	Eps
1	5	0.030864	0
2	5	0.030864	0
3	5	0.05	0

Table 107: SARS (y3) imputation statistics for LTILL

Experiment	alpha	beta	delta	Dcat	tj
1	0	0.007299	0.007299	0	0
2	0	0.007299	0.007299	0	0
3	0	0.007299	0.007299	0	0

Table 108: SARS (y3) error statistics for TENURE

Experiment	W	D	Eps
1	3.583561	1	1
2	3.583561	1	1
3	41.690141	1	1

Table 109: SARS (y3) imputation statistics for TENURE

Experiment	alpha	beta	delta	Dcat	tj
1	1	0.017355	0.032547	0.01546	0.017595
2	0.894737	0.024793	0.038242	0.013832	0.015743
3	0.894737	0.024793	0.038242	0.013832	0.015743

Table 110: SARS (y3) error statistics for HHSPTYPE

Experiment	W	D	Eps
1	56.363439	0.835294	0.747255
2	69.230785	0.770833	0.673116
3	5.580102	0.65625	0.509675

Table 111: SARS (y3) imputation statistics for HHSPTYPE

Experiment	alpha	beta	delta	Dcat	tj
1	0.865385	0.007679	0.044118	0.036765	0.038148
2	0.865385	0.007679	0.044118	0.036765	0.038148
3	0.865385	0.007679	0.044118	0.036765	0.038148

Table 112: SARS (y3) error statistics for ROOMSNUM

...

Experiment	W	D	Eps
1	41.678913	0.811765	0.717647
2	41.678913	0.811765	0.717647
3	5.580102	0.65625	0.509675

Table 113: SARS (y3) imputation statistics for ROOMSNUM

Experiment	alpha	beta	delta	Dcat	tj
1	1	0	0.070306	0.070306	1.173275
2	0.423529	0.010676	0.039702	0.029777	0.496916
3	0.423529	0.010676	0.039702	0.029777	0.496916

Table 114: SARS (y3) error statistics for BATH

Experiment	W	D	Eps
1	1	0.011905	0
2	1.522223	0.089655	0
3	1	0.011905	0

Table 115: SARS (y3) imputation statistics for BATH

Experiment	alpha	beta	delta	Dcat	tj
1	1	0	0.067406	0.067406	0.154811
2	1	0.081427	0.143345	0.067406	0.154811
3	1	0.081427	0.143345	0.067406	0.154811

Table 116: SARS (y3) error statistics for CENHEAT

Experiment	W	D	Eps
1	48.241379	0.487603	0.357454
2	132.95628	0.690476	0.613693
3	56	0.46281	0.329549

Table 117: SARS (y3) imputation statistics for CENHEAT

Experiment	alpha	beta	delta	Dcat	tj
1	0.724138	0.006103	0.040664	0.034855	0.564813
2	0.431034	0.011334	0.031535	0.020747	0.336198
3	0.431034	0.011334	0.031535	0.020747	0.336198

Table 118: SARS (y3) error statistics for INSIDEWC

Experiment	W	D	Eps
1	7	0.063063	0
2	13	0.097015	0
3	0	0	0

Table 119: SARS (y3) imputation statistics for INSIDEWC

SARS (y3) error statistics for CARS are all zero.

Experiment	W	D	Eps
1	0.847931	0.640351	0.528015
2	0.847931	0.640351	0.528015
3	0.847931	0.640351	0.528015

Table 120: SARS (y3) imputation statistics for CARS