WP. 21

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

WORK SESSION ON STATISTICAL DATA EDITING
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

USING MIXTURE MODELLING TO DEAL WITH UNITY MEASURE ERROR

Supporting Paper

Submitted by ISTAT, Italy[1]

I. INTRODUCTION

1. In the context of Official Statistics, one important quality aspect to deal with is data accuracy. We adopt the definition suggested in the Encyclopedia of Statistical Sciences, (1999): "accuracy concerns the agreement between statistics and target characteristics". A number of factors can cause inaccuracy along the overall statistical survey process. Data inaccuracy can be reduced during the Editing and Imputation phase (E&I), which can be viewed as a "data quality improvement tool by which erroneous or highly suspect data are found, and if necessary corrected (imputed)" (Federal Committee on Statistical Methodology, 1990).

2. The complexity of the investigated phenomena and the existence of several types of non-sampling errors make the E&I phase a very complex task. In the E&I literature a common error classification leads to define two different error typologies: *systematic errors* and *random errors*: the former relate to errors which go in the same direction and lead to a bias in statistics, while the latter refer to errors which spread randomly around zero and affect the variance of estimates (Encyclopedia of Statistical Sciences, 1999). Understanding the nature of errors is not only useful in order to identify their origin and to assess their effects on estimates, but also to choose the most appropriate methodology to deal with them (Di Zio et al., 2002). While the Fellegi-Holt approach (Fellegi *et al.*, 1976) is a well-established procedure to deal with random errors, systematic errors are generally treated through ad hoc solutions.

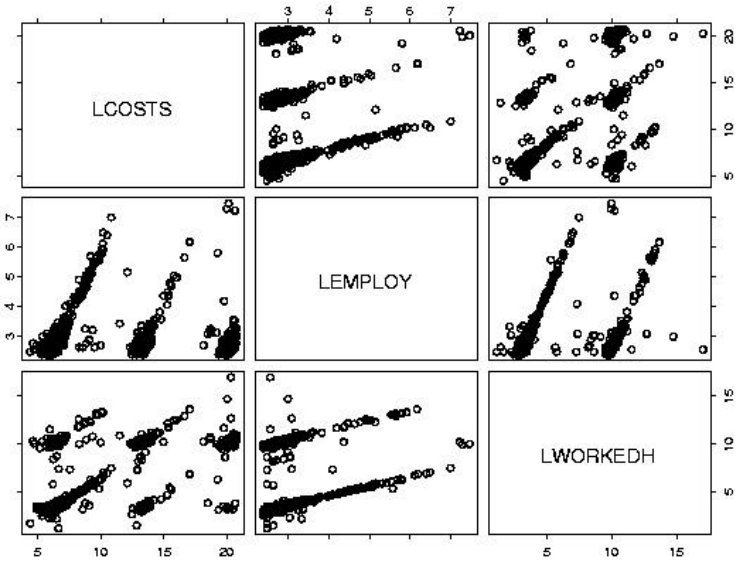
3. In the family of systematic errors, one that has a high impact on the final estimates and that frequently affects data in statistical surveys measuring quantitative characteristics (e.g. business surveys) is the *unity measure error times a constant factor* (e.g. 100 or 1,000). This error is due to the erroneous choice, by some respondents, of the unity measure in reporting the amount of questionnaire items.
As real examples of surveys affected by systematic errors due to unity measure errors, we selected two ISTAT investigations: the 1997 Italian *Labour Cost Survey* (LCS) and the 1999 Italian *Water Survey System* (WSS).

4. The LCS is a periodic sample survey that collects information on employment, worked hours, wages and salaries and labour cost on about 12,000 enterprises with more than 10 employees. In figure 1 the logarithmic transformations of *Total Labour Cost* (LCOST), *Number of Employees* (LEMPLOY), *Worked Hours* (LWORKEDH) are represented in a scatter plot matrix. Note that the employment variable at this

[1] Prepared by Marco Di Zio (dizio@istat.it), Ugo Guarnera (guarnera@istat.it), Orietta Luzi (luzi@istat.it)

editing stage is error free because of a preliminary check with respect to information from business registers (Cirianni et al., 2000). By considering the relation between *Total Labour Cost* and *Number of Employees*, the former results affected by two types of unity measure errors: by 1 million and by 1,000 factor. On the other hand, when also *Worked Hours* are taken into account, data show different combinations of these types of errors in *Total Labour Cost* and *Worked Hours* itself.

Figure 1: Multiple scatter plot between total labour cost, employees, worked hours (logarithmic scale)



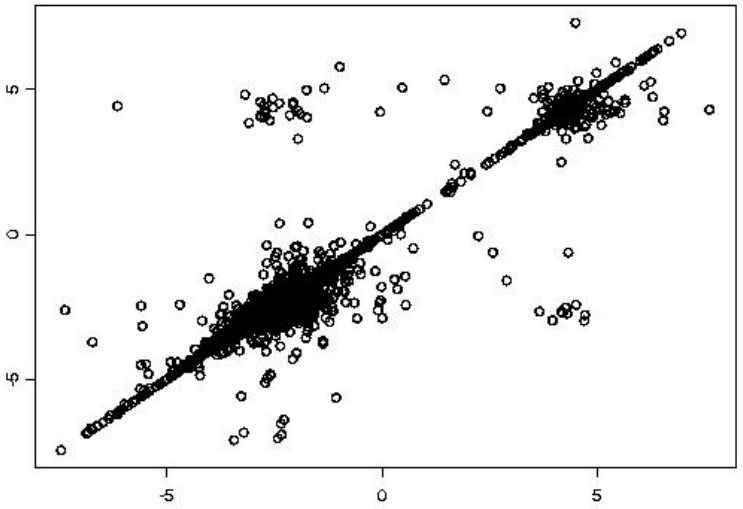
5. The *WSS* is a periodic total survey collecting information on water abstraction, supply and use on the 8,100 Italian municipalities. We restrict our analysis to two main variables measuring *Water Supplied* and *Water Invoiced*. Both these variables refer to total water volumes, and respondents are requested to provide all the quantities in thousands of cubic meters. During preliminary analyses performed on the per capita water quantities, we noticed the occurrence of the 1,000-factor unity error (figure 2). This is probably due to the misunderstanding of some respondents that expressed the water volumes in litres or in cubic meters rather than thousands of cubic meters, as requested.

6. Relating to this type of systematic errors, the critical point is the localisation of items in error rather than their treatment. In fact, once an item is classified as erroneous, the best action (treatment) is uniquely determined and consists in a deterministic action recovering the original value through an inverse action (e.g. division by 1,000) neutralising the error effect.

7. The unity measure error times a constant factor is generally tackled through ad hoc procedures using mainly graphical representations of marginal or bivariate distributions, and ratio edits. When adopting traditional approaches, the error localisation problem is not only complex, but also cost and resources consuming.

8. In terms of complexity, two elements are to be considered: identifying "true" errors is difficult when the actual variable values have a high variability that makes part of the real values distribution overlapping to part of the erroneous values distribution; furthermore, traditional techniques allow no more than pair-wise comparisons between variables, thus it is generally impossible to take into account multivariate relationships in the error localisation phase.

Figure 2: Multiple scatter plot between per capita water supplied (x axis) and per capita water invoiced (y axis) (logarithmic scale)



9. Relating to costs, the elements to be taken into account are both the complexity of designing and implementing ad hoc deterministic procedures aiming at automatically identifying such types of errors, with the associated risk of misclassifying certain amounts of units, and the resources spent in manually checking high amounts of observations having different probabilities of being in error, with the associated risk of over-editing.

10. In this paper we propose a probabilistic formalisation of the problem and an approach based on mixture modelling (McLachlan *et al.*, 1988). This approach leads to a rigorous formalisation of the problem allowing an estimate of the conditional probability that an observation is affected by this kind of error.
The main advantages of the mixture approach are that it allows a multivariate analysis instead of a pairwise analysis of data, and that it provides elements that can be used to optimise the trade off between the automatic and interactive approaches to the problem, i.e. between costs and accuracy.

11. In this work the proposed method is illustrated (sections 2 and 3) and an application on experimental data is described (section 4).

II. THE MODEL

12. It is hard to give a comprehensive formalisation of random and systematic errors. In this context, we provide a definition that, although not exhaustive, includes many common situations. Let the surveyed variable X (vector) be a random variable with $E(X)=\mu$ and $Var(X)=\sigma^2$, possibly affected by random and/or systematic errors. We assume that systematic error affects the variable only in its expected value, that is $E(X)=g(\mu)$, (generally an additive error mechanism is assumed so that $g(\mu) = \mu + constant$), on the other hand a random error mechanism has impact only on the Variance (Covariance) structure of the X variable such that $Var(X) = \sigma_e^2$ (generally $\sigma_e^2 >> \sigma^2$). Thus, the main consequence of the systematic error is that estimates are biased while random error reduces the precision of estimates.

13. We restrict our attention to systematic errors. Following the formalization so far introduced, our goal becomes to assign data to different groups, each characterised by a specific "error pattern". To this aim we draw experiences from the cluster analysis theory. Let us suppose we have n observations of a vector of q

variables $X^j = (X^j_1,\ldots, X^j_q)$ $(j=1,\ldots,n)$ iid with respect to a distribution $f(X_1,\ldots,X_q ; q)$, with $E(X_1,\ldots,X_q)=(m_1,\ldots,m_q)$ and $Var(X_1,\ldots,X_q) = S$.

14. We assume that systematic errors affect the random vector X only by transforming its expected value m into $g_i(m)$ where $g_i(\cdot)$: R^q? R^q $(i=1,\ldots,h)$ are a set of known functions. Thus, the functions g_i are associated with h distinct clusters, each characterised by a distribution having the same parametric form as the original one, but mean vector transformed according to the corresponding function g_i. For instance, if the sistematic error acts on all the variables X_s $(s=1,..q)$ in the same manner by transforming their expected values m_s according to: m_s ? $g^*(m_s)$ where g^* is a specified function, the number of clusters (error patterns) will be $h = 2^q$, i.e. the number of different combinations of error occurence on the q variables. In this case, each function g_i and each corresponding cluster, is associated with one of the 2^q possible sub-sets of variables affected by the error, e.g. the group characterised by the expected value $E(X)=(m_1, g^*(m_2),m_3,m_4,\ldots,m_q)$, is a cluster of units with errors affecting only the variable X_2. We remark that we assume a common covariance matrix because we make the hypothesis that the possible random error acts in the same way on all the data. For the error localisation purpose we follow a model-based approach based on finite mixture models, where each mixture component G_i (i=1,..h) represents a single error pattern. Formally, we assume that our sample $X^j=(X^j_1,\ldots,X^j_q)$ $(j=1,\ldots,n)$ is iid w.r.t $\sum_{t=1}^{h} p_t f_t(\cdot ; q_t)$, where $\sum_t p_t = 1$ and $p_t \geq 0$.

15. In order to classify an observation x^j in one of the h groups, we compute the posterior probability $t_j(x^j ; q, p) = pr(j\text{-th observation } \hat{I}\ G_i \mid x^j ; q, p)$, that is

$$t_i(x^j ; q, p) = p_i f_i(x^j ; q_i) \Big/ \sum_{t=1}^{h} p_t f_t(x^j ; q_t) \quad i=1,\ldots,h. \tag{1}$$

Thus, an observation is assigned to the cluster G_t, if

$$t_t(x^j ; q, p) > t_i(x^j ; q, p) \quad (i=1,\ldots,h ; i^1 t).$$

16. The previous allocation rule is the optimal solution for the classification problem, in the sense that it minimises the overall error rate (Anderson, 1984, Chapter 6).

17. Since the parameters (q_i , p_i) are unknown, we use the maximum likelihood estimates $(\hat{q}_i ,\hat{p}_i)$ in order to classify data, then the classification rule becomes:

$$t_t(x^j ; \hat{q}_t ,\hat{p}_t) > t_i(x^j ; \hat{q}_i ,\hat{p}_i) \quad (i=1,\ldots,h ; i^1 t). \tag{2}$$

18. We make the assumption that the $f_t(x ; q_t)$ is multivariate normal density $MN(m_t; S)$. The model assumed is suitable to deal with the unity measure problem. As matter of fact, in business surveys, variables are frequently considered log-normal with a multiplicative factor error. Thus in many applications we are legitimate to use the previous modelling by adopting logarithmic transformations on data.

19. In order to compute the likelihood estimates, we use the EM algorithm as suggested in McLachlan *et al.* (1988). Nevertheless, an additional effort was needed to adapt the algorithm to our particular situation, where the mean vectors of the mixture components are linked by a known functional relationship.

III. DIAGNOSTICS FOR SELECTIVE EDITING

20. Once the parameters of the mixture have been estimated, we are able to classify data into the different clusters, in other words for each observation we can assess whether it is in error or not, and in which variable the error is present. However, in general not all the observations can be classified with a high degree of belief, and some other units can be “atypical” with respect to the estimated model. In other words, different types of critical observations can be identified after the modelling phase: those units that have been classified in a cluster, but at the same time have a not negligible probability of belonging to another cluster, and those observations that are outliers with respect to the model so far introduced. Being these observations

potentially affected by an error in one or more variables, in order to increase the data quality it would be useful to make on them a double check (through a clerical review or, in the most difficult cases, with a follow-up). However, the need of reducing possible over-editing and the interactive editing costs, the manual review and/or follow up should be concentrated on the "most critical" observations in terms of either probability of belonging to a cluster or atypicality. To this aim, diagnostics directly provided by the proposed mixture model can be used.

21. In order to measure the degree of belief in the class assigned to an observation, we can consider the probability (2). When this probability is not very different compared to its complement, we have that the observation has a not negligible probability to belong to another cluster. These observations are those in the region where the mixture components are overlapping.

22. In addition to these units, we have also observations that are far from all the clusters (all the mixture components), i.e. the outliers with respect to the model introduced. Also these observations can hide an error. One way of approaching the problem is described in sec. 2.7 of McLachlan (1988).

23. Let x_i^j for $j = 1,...,\hat{m}_i$ be the observations assigned to the *i*-th cluster ($i = 1,...,h$) where $\hat{a}_{ij}$ is the corresponding areas to the right of the following value

$$\frac{(n\hat{m}_i / q)D(x_i^j; \hat{m}_i, \hat{\Sigma})}{(n+q)(\hat{m}_i - 1) - \hat{m}_i D(x_i^j; \hat{m}_i, \hat{\Sigma})} \qquad (3)$$

under the $F_{q,n}$ distribution, $D(\cdot,\cdot)$ is the Mahalanobis squared distance, and $n = n - h - q$. Under the normality hypothesis, the lower is $\hat{a}_{ij}$ the higher is the probability of x_i^j of being atypical for our model.

24. Thus both the classification probabilities and the atypicality indicator $\hat{a}_{ij}$ can be viewed as diagnostics useful in order to prioritise observations to be accurately investigated for assessing whether or not they are affected by errors. The classification probabilities and the atypicality index could be used, according to a *selective editing/significance editing* approach (Latouche *et al.*, 1992, Lawrence *et al.*, 2000), to build up appropriate score functions to prioritise critical units.

IV. AN ILLUSTRATIVE EXPERIMENTAL EXAMPLE

25. In order to show the advantages of the illustrated approach, we have performed the following example. We considered three variables (*X, Y,* and Z), and we simulated the case when all the variables are corrupted by a systematic error that is the addition of a log(1,000) factor.

26. A sample of 1,000 observations from a mixture of eight multinormal, i.e. $\sum_{t=1}^{8} p_t MN_t(\cdot; m_t, \Sigma)$ were generated, where π=(0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.2), m_t is obtained by adding log(1,000) to all the possible combinations of the components of the vector $m = (m_x, m_y, m_z) = (-2.554292, -2.533458, -2.224171)$ (see table 2 expressing the eight components), and the covariance matrix Σ is reported in Table 1[2].

Table 1: Covariance matrix used for simulating the mixture

	X	*Y*	*Z*
X	2.499156	2.162238	2.188616
Y	2.162238	2.242727	2.002701
Z	2.188616	2.002701	2.718579

[2] These parameters have been estimated on WSS

27. Under this setting, we used the mixture model approach described in sections 2 and 3 on (*X,Y,Z*), and classified all data in one of the eight possible clusters. The application was performed by using a generalised procedure, developed in the SPSS programming language R, that allows to both perform the estimation and classification steps, and compute the final diagnostics (classification probabilities and atypicality index) needed for selecting critical units.

28. In order to mime the traditional situation, we have compared the results obtained with the mixture multivariate approach with the results obtained following a pairwise classification on all the couples coming from (*X,Y,Z*). In other words, we considered the eight clusters (mixture components) representing the error composition, reported in the following table 2.

Table 2: Mixture components on experimental data

Cluster	*Error location*
(0,0,0)	No errors
(0,0,1)	Error in Z
(0,1,0)	Error in Y
(1,0,0)	Error in X
(0,1,1)	Error in (Y,Z)
(1,0,1)	Error in (X,Z)
(1,1,0)	Error in (X,Y)
(1,1,1)	Error in (X,Y,Z)

29. While in the multivariate approach the classification is directly performed through the mixture modelling, in order to obtain a pairwise classification we need to firstly define the four clusters on the couples *(X,Y), (X,Z), (Y,Z)* listed in the following table 3.

Table 3: Possible cluster for each couple of items

Cluster	*Error location*
(0,0)	No errors
(0,1)	Error in the 2nd variable
(1,0)	Error in the 1st variable
(1,1)	Error in both variables

30. According to the pairwise classification, we obtained the final classification in the eight clusters through the scheme reported in the following table 4.

Table 4: Possible cluster combinations and final classification for triplet of items

Cluster (X,Y)	*Cluster (X,Z)*	*Cluster (Y,Z)*	*Cluster (X,Y,Z)*
(0,0)	(0,0)	(0,0)	(0,0,0)
(0,0)	(0,1)	(0,1)	(0,0,1)
(0,1)	(0,0)	(1,0)	(0,1,0)
(1,0)	(1,0)	(0,0)	(1,0,0)
(0,1)	(0,1)	(1,1)	(0,1,1)
(1,0)	(1,1)	(0,1)	(1,0,1)
(1,1)	(1,0)	(1,0)	(1,1,0)
(1,1)	(1,1)	(1,1)	(1,1,1)

31. Note that the list reported in table 4 is not comprehensive of all the possible pairwise combinations. All the not considered combinations are incoherent, e.g. the combination (1,0),(0,0),(0,0) suggests that *X* is both incorrect (from the first couple) and correct (from the second couple), two situations clearly

incompatible. Thus, we need an additional cluster (cluster 9), containing all the incompatible pairwise conclusions.

32. In table 5 the application results in terms of final classification under the pairwise (*ClxyzBiv*) and the mixture (*ClxyzMixt*) methods are shown, for the subset of units either not classified or erroneously classified by the two approaches, The true cluster each unit belong to is indicated in the column *Cltrue*. The last two columns contain the classification errors under the pairwise (*ErrBiv*) and the mixture (*ErrMixt*) approaches.

Table 5: True clusters, final classifications under the pairwise and model based approaches, classification errors for

X	*Y*	*Z*	*Cltrue*	*Clxy*	*Clxz*	*Clyz*	*ClxyzBiv*	*ClxyzMixt*	*ErrBiv*[3]	*ErrMixt*[4]
1,01	0,03	1,55	1	1	4	1	9	1	*nc*	0
2,24	2,44	3,03	1	4	4	4	8	8	1	1
0,44	1,70	1,40	1	4	1	4	9	8	*nc*	1
1,02	2,24	7,84	2	4	2	2	9	2	*nc*	0
1,06	7,87	1,64	3	2	4	3	9	3	*nc*	0
1,19	7,55	2,18	3	2	4	3	9	3	*nc*	0
7,11	1,21	1,50	4	3	3	4	9	4	*nc*	0
-5,86	0,52	2,07	5	2	2	1	9	5	*nc*	0
0,55	0,61	0,52	8	1	1	1	1	1	1	1
1,31	0,66	1,37	8	1	4	1	9	1	*nc*	1
-0,26	0,77	1,28	8	1	1	1	1	1	1	1
1,56	1,03	1,05	8	4	4	1	9	8	*nc*	0
0,94	-0,17	0,13	8	1	1	1	1	1	1	1
1,43	2,63	-0,68	8	4	1	3	9	8	*nc*	0

33. From these results is clear that the multivariate approach is very useful to disentangle the situations that that cannot be classified by using a pairwise approach (shadowed cells). From table 5 we also note that there are a certain number of observations that are misclassified by the multivariate approach too. The misclassification relates to clusters 1 and 8, corresponding to cases in which all variables are either correct or erroneous. In practice, these units are those in the overlapping region among the component mixtures, and only through either a re-contact or the use of auxiliary information they can be correctly classified and treated. This is why it is important to use the diagnostics provided by the mixture model to prioritise the observations to be re-contacted.

34. In order to understand which units in the overlapping region are more important in terms of doubt assignment to a given cluster, we could use for example a ranking based on the probability of being classified in a different cluster. By ranking units in descending order with respect to this probability, and selecting the first 1% of them, it results that out of the six units misclassified by the model, four are pointed out for re-contacting (they are in fact in the first seven positions, see table 6).

35. Finally, we have also considered the atypicality index (table 7). We note that with this index we find the first unit misclassified at the 42nd position. This is mainly due to the fact that data have been simulated exactly from a multinormal distribution. Its usefulness is expected particularly in real cases.

36. In order to better understand which units are pointed out by these two different scorings, we report the scatter plot matrix (figure 3) where the observations selected by the score based on the classification probability (red stars), and the observations selected by the atypicality index (blue triangle) are simultaneously depicted.

[3] *nc*= not classified unit; 1=unit erroneously classified; 0=unit correctly classified.
[4] 1=unit erroneously classified; 0=unit correctly classified.

Table 6: 1% selected units according to the classification probability

X	*Y*	*Z*	*Cltrue*	*ClxyzMixt*	*1-Class. prob*
1,56	1,03	1,05	8	8	1-0,57
1,31	0,66	1,37	8	1	1-0,59
0,44	1,70	1,40	1	8	1-0,70
0,24	0,97	0,08	1	1	1-0,79
0,94	0,97	-0,84	1	1	1-0,79
-0,26	0,77	1,28	8	1	1-0,80
0,55	0,61	0,52	8	1	1-0,82
0,76	1,65	2,37	8	8	1-0,82
1,01	0,03	1,55	1	1	1-0,83
1,02	1,83	1,71	8	8	1-0,84

Table 7: 1% selected units according to the atypicality index

X	*Y*	*Z*	*Cltrue*	*ClxyzMixt*	*Atypic*
1,43	2,63	-0,68	8	8	0,0007
1,94	4,43	4,26	8	8	0,0011
9,37	7,59	1,88	7	7	0,0026
7,96	9,43	2,03	7	7	0,0031
-5,56	-5,44	-2,48	1	1	0,0073
1,02	2,24	7,84	2	2	0,0084
-7,55	-7,41	-0,07	2	2	0,0108
1,12	-3,94	-3,72	4	4	0,0121
3,09	2,15	-6,01	7	7	0,0133
4,58	5,48	2,58	8	8	0,0147

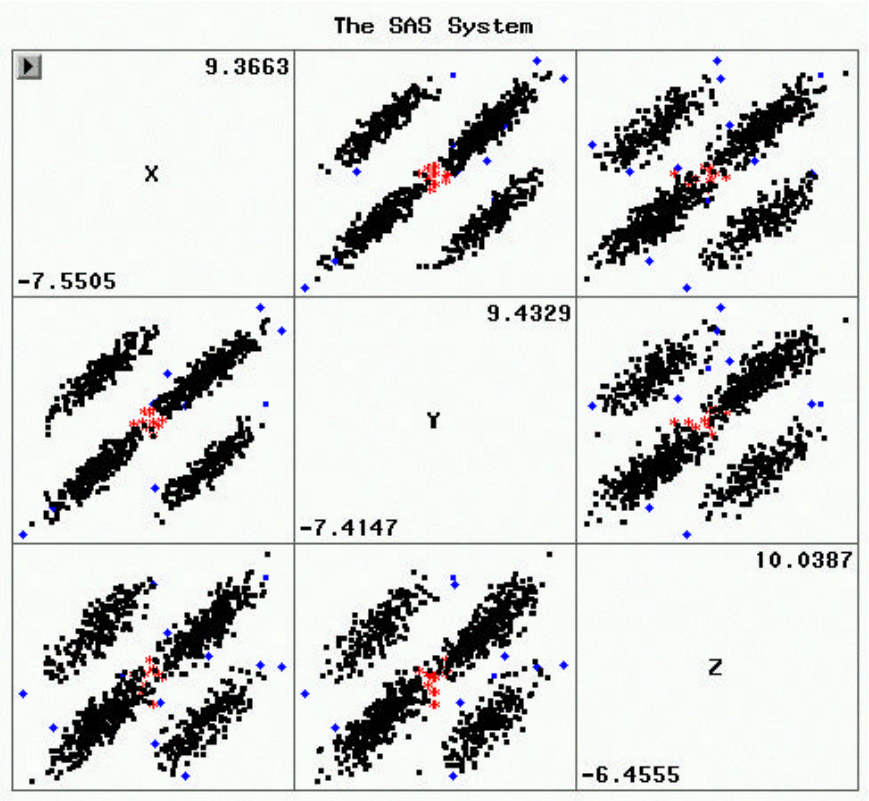
V. FINAL REMARKS AND FURTHER RESEARCH

37. In this paper we propose a mixture model to deal with the unity measure error times a constant factor that frequently affects numerical continuous survey data (in particular, business data). This approach has the advantage, with respect to traditional ones, to formally state the problem in a multivariate context, providing also a number of useful diagnostics for prioritising doubtful units possibly containing potentially influential errors. Therefore, the method leads us to a less subjective approach to the problem, giving to subject matter experts the possibility of dealing with the problem using a generalised approach less cost and time consuming than the traditional ones.

38. The potential benefits that can be obtained by using this approach, in terms of both capability of correctly classifying units depending on the error patterns, and usefulness in identifying doubtful units on which manual reviews/follow up activities are to be concentrated, have been showed through an experimental example. Furthermore, this approach can be easily implemented in generalised software that from one hand allows to automatically resolve a high percentage of erroneous cases, and from the other hand automatically provides useful elements to optimise the interactive data review.

39. However, this work is just a first experiment, more studies are needed. In particular, we feel that further analyses should be devoted to the use of distributions other than multinormal, and also a non-parametric approach. Interesting developments relate to the use of the model diagnostics in the context of selective editing score functions explicitly designed for the efficient localisation of this type of errors.

Figure 3: Scatter plot matrix for the three simulated items



REFERENCES

Anderson, T.W. (1984). *An introduction to Multivariate Statistical Analysis*. Second Edition. New York: Wiley.

Encyclopedia of Statistical Sciences (1999). Update Volume 3, pp 621-629. J. Wiley & Sons, Inc.

Cirianni A., Di Zio M., Luzi O., Seeber A.C. (2000). "The new integrated data editing procedure for the Italian Labour Cost survey: measuring the effects on data of combined techniques", *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, June 7-21.

Di Zio M., Luzi O. (2002). "Combining methodologies in a data editing procedure: an experiment on the survey of Balance Sheets of Agricultural Firms", *Statistica Applicata*, **14**, N.1.

Federal Committee on Statistical Methodology, (1990). *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18.

Fellegi I.P., Holt D. (1976). "A systematic approach to edit and imputation", *Journal of the American Statistical Association*, **71**, pp 17-35.

Latouche, M., Berthelot, J.M. (1992). "Use of a Score Function to Prioritise and Limit Recontacts in Business Surveys", *Journal of Official Statistics*, **8**, pp. 389-400.

Lawrence, D., McKenzie, R. (2000). "The General Application of Significance Editing", *Journal of Official Statistics*, **16**, pp. 243-253.

McLachlan G. J., Basford K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.