NATIONS UNIES | ОБЪЕДИНЕННЫЕ НАЦИИ | UNITED NATIONS

COMMISSION ECONOMIQUE POUR L'EUROPE | ЭКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ | ECONOMIC COMMISSION FOR EUROPE

SEMINAIRE | СЕМИНАР | SEMINAR

STATISTICAL COMMISSION AND
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN
STATISTICIANS

Distr.
GENERAL

CES/SEM.38/27
17 March 1998

ENGLISH ONLY

Seminar on integrated statistical information systems and related matters (ISIS '98)
(Bratislava, Slovak Republic, 26-29 May 1998)

Topic (iv): automation of data collection and its relation with the achievement of lower costs and greater efficiency

USE OF EXPLORATORY DATA ANALYSIS AND DATA MINING FOR STATISTICS

Submitted by Statistics Netherlands[1]

I. INTRODUCTION

1. In the last few years much has been publicized about Data Mining (DM), and expectations have been raised about the new possibilties to discover ‘hidden, valuable information’ in large data bases.[FAY96, ACM96, HEC97] As statistical offices have usually many large data sets from various sources, it is important to asess the potential of DM for the re-use of all these data sets. Not only could DM perhaps help in finding interesting and reliable findings in the data that are processed already for many years, but perhaps DM could also help to reduce the number of separate data collections that we now maintain. In this contribution an attempt is made to sketch what we can expect in the near future from DM, by comparing it with other, more traditional forms of Exploratory Data Analysis.

II. WHAT IS EDA

2. Exploratory Data Analysis (EDA) predates the computer era.[TUK77] It is data analysis with an unfocussed goal, and usually in an early stage. The goal is the discovery of hitherto unknown properties of the data set, especially relations between variables. Usually this implies that the data set is ‘new’ to the analyst, though there is sometimes occasion to look at a well known data set from a fresh point of view. Even if the data set is ‘new’ to the analyst, usually some background knowledge and *a priori* assumptions about the data are needed. Even though a *small* data set can be analysed if there comes no documentation at all with the data [ISR82], it is very hard indeed to analyse a data set with many variables when it is unknown what the variables stand for.

3. The discovery of unknown properties and relations is at the one end of a continuum, the other end of which is formed by the fine tuning of an established relation or the improvement of the degree of certainty of an accepted model. In the detection of unknown properties, it is not the intention to ‘just discover

[1] Prepared by J.W.P.F. Kardaun and J.G. Bethlehem.

something new'. Quite often some mental image of the topic under study is present, and some — quite implicit — expectations about how the data should be like are available. But the analyst is looking in a less focussed way at the data set, both in order to find 'real' properties, and to discover how the data are actually collected and represented. It is a well known fact to analysts, that the metadata[2] to data sets are quite often idealized.[3] Moreover, in the early stages of analysis, the distributions of the variables, the missing value pattern, the main primary relations between the variables, and the structure of the data set is 'learned' by the analyst. Once an analyst is familiar with a data set, he can choose one or several suitable analytic methods (see section V. B. Interpretation), of course also with an eye towards some research question. In this early stage also suitable transformations to variables (e.g. logarithmic, logistic, discretisation) are performed, aimed at making the representation of the data better match the analytic methods.[4]

4. Here EDA more or less ends, as further steps in the analysis evolve to be more in-depth-analytic. Is there a sharp distinction between EDA and in-depth-analysis? Probably not.[5] Not so, because there is a continuous transition between vague, widely and narrowly-focussed analytic goals, and consequently instruments. And in all stages of analysis, one has to keep an open mind for pitfalls of artefacts, errors, bias, noise, and misinterpretation. Probably, because it may be justified to require in the EDA phase that statistical testing of hypotheses is *not* performed.[6] Moreover, the boundary between EDA and in-depth-analysis is becoming even more blurred by the development of fast computers and sophisticated software. In a modern sense of the word, EDA may be considered to pertain to a family of multiple runs of, say, discriminant analysis with a slightly differently defined target variable; or to make thousands of cross tabulations in search of a 'hit'. The word *exploratory* in EDA has had also the connotation of *easy to perform,* as it is wise to do the simple things before the hard ones. But what is easy to perform, changes with technological developments.

III. WHAT IS DM

5. Data mining is a combination of computational and statistical techniques to perform EDA on rather large, and mostly not very well cleaned data sets (or data

[2] Meta-data means data about the data, not only record description and variable codings, but also how the data were collected and what their 'real life' meaning is.

[3] It is a sad fact of life, that most data sets, even the well documented ones, can not fully be understood from the meta-data or documentation. Quite often it is necessary to call upon the data collectors and documenters to clear up important aspects.

[4] From a (mathematical) statistician's point of view, it can hardly hurt to perform transformations that gives more *normal* distributions. From an analyst's point of view, this may not always be helpful, as the transformed variables may have no longer any 'real life' meaning.

[5] Purists would say: "There *should* be." From a didactical and theoretical point of view, it is good to make a sharp distinction between (a) data cleaning, (b) exploration, (c) formulation (but not testing of) hypotheses, and (d) testing of hypotheses. (a)–(c) are considered EDA. But our point here is that these steps are often interleaved, not only because of practical reasons, but also because data cleaning, exploration and formulation each influence each other.

[6] If EDA is limited to studying one data set and making no inferences, i.e. no generalisations towards larger populations, then statistical testing is often said to make no sense, as the data set is 'the whole universe'. But, even if a data set is the whole universe, it is useful to how homogeneous the data set is for a given finding, which requires testing. A high homogeneity gives more confidence about future findings, it is a kind of internal generalisation.

bases).[7] The term is a kind of warrior's name, as data miners are well aware of the fact that 'pure' statisticians dislike the activity of *looking for significance,* which they also call derogatorily *data fishing* or *data dredging*. Data miners also know that the metaphore of mining as a sturdy, industrial and profitable process is attractive to many other types of persons, especially if the mining is considered to be gold mining. Be that as it may, data mining is usually more applied (and so borrows from disciplines such as *decision support systems, management information systems*) than EDA, which grows from the tree of statistics. Moreover, data mining is heavily dependent upon data base theory and techniques and on artificial intelligence achievements for the realization of it's promises. The ultimate *raison d'être* of data mining is: "However fast computers will become, there will always be data sets that are so large and have so many variables, that an exhaustive search for all potential interesting findings is impossible. Therefore, we need (computerized) techniques that make a good guess in selecting most of the areas that are most promising to scrutinize further." Especially in situations when there is little 'hard' *a priori* knowledge, or when we expect highly non-linear effects or complex interacions, or when deep insight is less important than simulation or taking appropriate measures, data mining can be more useful than real statistical analysis.

IV. ROLE OF EDA IN STATISTICAL OFFICES

6. Acknowledging the fact that there is not one role model for statistical offices, it can be said that they have some things in common. One of those is that (re)production has more emphasis than new design or entering undiscovered grounds. Of course, the ongoing developments incur a need for innovation and research, but statistical offices have a tendency to do research only in directions that can later repeatedly be applied for 'production'. A single instance analysis of a data set is not unknown in statistical offices, but in most offices it will be relatively rare. National statistical offices that are leaning more towards a consultancy role in society, as can be found nowadays, might have more use for *one-off* analyses. Yet, EDA plays a role in statistical offices, for the purposes of (not in order of importance):

- feeling familiar with the data, and knowing your data thoroughly;
- general quality control, data cleaning, or more specific error searching when strange things are coming out
- detecting the most important new or newsworthy features of a dataset
- special analyses, perhaps paid for by third parties the development of new statistics
- creating 'new' data sets by combining or matching 'old' data sets.

All together, this means that EDA is always a step that is part of a larger process, and that EDA has a limited role in statistical offices. Still, EDA has been performed daily in statistical offices, and some form of (mostly informal) standard practice exists. This will be described now.

V. EDA BY HAND

7. EDA 'by hand' does not mean using pencil and paper only, but without the use of typical data mining software. Traditional software tools, such as aggregating, tabulating, (plain) displaying software and the 'classical' statistical packages (SPSS, SAS) can be used to do EDA, where every (small) step is determined by human intervention and preparation (hence: 'by hand'). As mentioned briefly in the introduction, there are several levels of unfamiliarity with a data set, which for

[7] In this paper we will not cover the relation of DM with *Data Warehousing*, i.e. taking care that data from different sources are neatly organised, can be linked to each other and use the same definitions for (essentially) the same phenomena. We do not require that DM is acting on data from different sources.

the purpose of this treatise we will distinguish in *representation* and *interpretation* — even though these elements are related.

A. Representation

8. If a data set is completely new or unknown, the first task is to study the format of the data set, the records and the variables. Based upon a combination of the documentation — often obscure — and some trial tables, for each value of each variable it has to become clear what it stands for. Special attention is needed for structured data (such as sub-questionaires, repeated questions) as a notation for "N/A" (not applicable) on the single variable level often does not reflect the reason why this (block of) variable(s) is skipped. Similarly, a very tricky matter is always the notation for the various reasons of "missing" data, which sometimes can be coded unexpectedly as 31 June (for a missing date), or as 99999 or –1 (for a missing number). Also, some effort is needed to find out that every variable is multiplied with the appropriate weight, if weighted sampling has been used. Yet another obligatory task: find out whether some data have been imputed, and how these can be recognised.

9. This step is often omitted from textbooks, as the examples in there are usually based upon small and easy to grab data sets. With large data sets there is need for a round of preliminary contact with the data set, before looking at individual variables.

10. The next step is usually a univariate check-up. The distribution of the variables is surveyed, and some transformation or recoding variables is done, e.g. a log transform to make the distribution less skewed, a discretisation (taking care that this does not hide peaks or introduce spurious effects), and some combinations of values. During this process, also a preliminary per - variable removal (or setting to a reasonable limit) of outliers is performed, usually rather intuitively.

11. A next step is often multivariate. It consists of inspection of two- or three-way[8] combinations of the most important variables, either as scatterplot,[9] as cross table or as correlation matrices. This frequently gives information about not-documented, but already elsewhere well-known, properties of the data collection.

12. Even if the data set is obtained from a very reliable source and contains no errors and comes with clear and detailed documentation, most of the above steps will have to be (quickly) performed in order to familiarise the analyst with the data.

13. Until now, the steps are not in considerable measure dependent upon the goal of any further analysis, called here the *research question*. Some of the steps could have been automated, but the meta-information about a data set is not always sufficient or not sufficiently processable in computer form. Moreover, if these steps are automated, there is no learning process in the mind of the analyst. Still, there is a need and plenty of opportunities for automating this, rather descriptive, part of EDA.

B. Interpretation

14. Gradually, the EDA proceeds to try to understand and perhaps model the relations between some of the variables. Now the steps are less standardized. One has to at least choose independent variables, dependent variables, some model or

[8] Usually not more than three, as higher dimensional tables are hard to oversee for humans.

[9] For possibilities of visual exploration of data, see [CLE94]

analytic technique. Quite often the complexity of the model (linear, higher order, number of interaction levels allowed) is reluctantly constrained, because of limitations in computing resources or available data. An important guide in this stage is the R^2 or similar measure[10] and the distribution of the residuals, which indicates how much we have gained or can gain by making the model more complex.

15. Sometimes the transformations mentioned in the representation section V.A. are motivated by the analytic method to be used or by intermediate results. For example, a discretisation of a dependent variable, if discriminant analysis is planned; or the combination of a few values of a discrete variable that have almost no different effect, if sparse cells are to be removed.

16. What analytical technique is used depends very much upon the research question – but also upon the available data and their type. Have we fixed the dependent and independent variables, or do we want to find good predicting (independent) variables? Have we fixed the order of the model-equation (linear, quadratic, higher) and the number and level of interaction terms and constants, or do we want to fit any model? All this can be considered part of EDA. If we have fixed the model, we may want to estimate the range of some parameters, if we think to know every (important) variable we may want to test some explicit (and perhaps intricate) hypotheses. The last two aims may not be EDA. But there is a relation between how we can fit models and test hypotheses: we need (almost) the same statistical approach (and software). This means that we have to think already at the EDA phase about the possibilities for later hypothesis testing.

C. Refinement

17. At some stage, the model is more or less clear, and some attempts are made, to make the model more robust or the effect more pronounced. This is not considered part of EDA anymore, as is verification of the findings on a *new* data set. Note that the ‘classical’ distinction of a descriptive step, followed by a generated hypothesis and statistical testing of the latter is neither completely followed, nor abandoned. In practice today, because of computing power, the descriptive step can sometimes be a description, list, of candidate hypotheses with test figures. Of course the role of ‘testing for 5% significance’ is changed considerably by this.

VI. DM AS EDA

18. Data Mining usually is not concerned with the very first steps of the ‘by hand’ phase: the data are supposed to be in a neat, well understood format, and their meaning should be clear.[11] If DM is supposed to run straight from (production) data bases, however, the different perspective that is applied to the data may well involve additional surveys of the type described in section V. A. Representation.

19. Currently, most data mining approaches make a clear distintion between the (calculation of the) criterion that has to be maximized (or minimized), e.g. an association measure or a distance measure, and the maximizing algorithm, often called the *search engine*.

[10] Provided that the case/variable ratio is high enough.

[11] DM is used by us in a limited sense. Others sometimes use the term for the whole process of cleaning, inspecting, exploring data and using the findings. For this we prefer the term KDD, which stands for Knowledge Discovery and Data Mining

A. Search engines

20. Given a one-dimensional criterion that is determined by a (very) high dimensional space of explanatory variables, with no *a priori* constraints on the irregularity of the surface, there exists no method besides exhaustive search to find with certainty the absolute maximum. If an exhaustive search is too expensive or too slow, there are various methods to reach with a reasonable likelihood a maximum that is more or less a global maximum. In essence, this goes as follows: from a starting point a local maximum is sought, i.e. following the gradient by some discrete steps, hoping that the step size is not so large that the gradient has reversed signs in between. In order to verify that not a very local maximum is reached, either some tolerance is built in of accepting small declines while continuing the search for a maximum, or a few values on some distance from the first found maximum are compared, or some alternative directions are tried for a few steps. Well known algorithms are, e.g. simulated annealing, tabu search, or genetic algorithms.[AAR97] Then new starting points, at a considerable distance from the first one and from the found local maxima are tried, to see whether they end on the same local maximum, or a lower maximum or a higher one. The choice of the first and the subsequent starting points is unimportant if the surface is rather smooth – fortunately this happens more often than we would dare to hope – but rather critical with irregular surfaces. Moreover, the speed with which a solution is found is very much influenced by a ‘lucky’ choice of starting values. Several approaches of generating starting points are in use, varying from the rigid equi-distant grids, to more dense samples in regions where the gradients are more varying, to pseudo-random points, or ‘promising points’ (genetic algorithms). All in all, the different search methods can be better (or worse) in two different aspects: is the maximum found really the global maximum and how fast is the result found.

B. Optimization criterion

21. As mentioned above, the criterion that is maximized has to be one dimensional. This is necessary for the search engines, and also because it is conceptually easier to think of maximizing in one dimension. As soon as maximization has to be performed on more than one criterion at a time, trade-off’s have to be specified if the gradient of some of them have a different sign. The specification of a trade-off function in general is quite difficult, and it mostly easier to make a choice afterwards, if the surface of the several single criteria is known. In this case simply looking for the maximum is not sufficient, one needs to have at least an impression of the surfaces of all criteria in the region between the individual maxima.

22. Constraining to single dimensional criteria: we can focuss our attention on measures of distance or closeness of groups, or of separation (as in discriminant analysis and classification), on similarity in time, on typicality (being “in the middle’’) or on atypicality (outliers), on change, on relatedness or indepedence, etc. The measure that statisticians are most used to fall in the catagories of association between variables, distance between groups and information (residues) of data sets.

VII. PITFALLS

A. Noise vs bias

23. When *large* data sets are analysed, the importance of random (or stochastic) noise becomes less important. With sufficient numbers virtually any difference, even negligible ones, become ‘significant’. So there is a false feeling that if one has large data sets available, it it easier to discover the ‘true’ differences. But, unfortunately, in practice there is another problem: most large data sets are less consistently collected. Differences in (space or time) local adminstrative

procedures (including the rigorousness of error checking and updating) and of definitions and notations, sampling and non-response make that many effects creeps into large data sets that can be easily confused with some ‘real’ effect. Scientific data collections often spend an extraordinary effort in excluding the above mentioned reasons for artefacts. Furthermore, scientific data collections often include *randomized* subgroups to cancel out the remaining effects of these and unknown artefacts. The price that these scientific data collections have to pay is mostly limited size. But the limitations that production type large data sets have, is that artefacts, biases, will be hard to exclude afterwards.

B. Linear models may become now non-linear

24. There is one more benefit of large scale data sets: finally we have sufficient data to try to fit more than a linear model with only a few interactions. In the past, with limited data sets, often linear or other simple relation between variables had to be assumed, because there would never be sufficient data to support anything fancier. How far can we go in adding terms and interactions? Is a model with 20 terms still useful for interpretation? If the size of the data set does not put a constraint on the class of the model, where do we have to stop? Clearly the answer lies outside traditional EDA.

C. Overfitting

25. Overfitting is making a model that performs well on one data set, but not on a ‘similar’ one (next year’s, e.g.). Basically, there is no remedy against overfitting, except for trying it on another data set. In practice, however, some assumptions about the representativity of the data set are made, supported by some other assumptions about the stochastic noise present and some limits to the interdependence of the variables. This prevents many cases of overfitting. More concretely, in many classical approaches of fitting, the addition of an extra term should give ‘sufficient’ reduction in residual variance (or deviance). This is not a final answer, as there are quite some assumptions underlying this approach, which are with practical data sets almost always violated (slightly).

26. A good thing about large data sets is, that they can be split up to see how models, fitted on one part, are performing on other parts. As most data sets that are used in data mining examples are ‘one of a kind’, that may be the only solution to make a statement about the (internal) generalisability of a fitted model. We are talking about bootstrapping and jack-knifing. Perhaps the worst thing that can be done with a large data set, is data mining on the whole data set.[12] It is likely not an important effect if it cannot be shown in a few thousand *relevant* cases. But often one need a data set of several hundred thousand cases, to keep a few thousand *relevant* cases after preprocessing. (For example: subgroups, interactions, boundary cells, etc.) If taking samples from the data set (for bootstrapping) is not done because it may dilute a supposed effect, we are in most cases not talking about large data sets anymore, and the DM/EDA aspects are giving way to the refinement stages of analysis.

D. First peeking ahead, then predicting the result

27. Data mining, but also current daily practice of EDA ‘by hand’ involves many cycles of generating still more interesting tables and ‘massaging’ the data. This is why data mining, now called data dredging, has such a bad name with many statisticians. What meaning has it that a certain model has a *p*-value of, say, 2.03%, if we just learned that a similar model had a *p*-value of 2.12%? The 5%

[12] Of course one can take the opposite approach, and check always whether an effect found in the whole data set is present in all parts of it; but human weakness makes that often the analysis stops if an effect is found in the whole data set.

significance test is too soft because of the repeated tests, and Bonferoni's correction is too harsh to be popular. Perhaps we should not look at significance while maximizing, but only once, afterwards, and use a statistical model for the signficance for finding a *maximized* effect. [MIL81].

28. Other areas which cause often problems with data analysis, such as the consistency between data sets, the definition of similar variables, and the treatment of the variable *time* will not be touched here, as they are not caused by DM nor solved by it. Still, they put constraints on the applicability of DM.

VIII. LIMITATIONS

29. As with any developing discipline, DM leaves currently many wishes open. So long as these are not yet fulfilled, applications that can be data mined on, have to be carefully selected. Some of these imperfections are also unresolved in EDA 'by hand' but there they can be more easily run around. Currently DM algorithms are notably weak in the handling of missing values, structured data, [13] (i.e. some blocks skipped, multiple answers) and repeated measurements, longitudinal data or time series. This means that 'real life' data sets, even after clean up, must be considerably preprocessed, before data mining can start. Moreover, until optimization criteria are developped that 'know' about these properties of data sets, the search engines are of limited use.

30. The data mining community has placed major emphasis on the use of DM for finding deviating subgroups, which is especially attractive for marketing purposes. Acknowledging that this is only one element of DM, it must be said that it is often done by looking at higher order interactions without looking at the main effects and lower order interactions. Without the notion of an (wisely computed) expected value, the extremes of the observed values have little to say. In other words, some work has to be done to reconcile (or augment) statistical theory with the new possibilties of automated searches.

IX. EXAMPLES

31. Well-known examples come from banking (credit card fraud detection), insurance (risk profiles, suspiciuous claims), retail trade ('market basket' analysis), etc. In statistical offices there are not so many examples of completed DM projects. Therefore, a few examples of potential questions that could be studied by statistical offices, and where DM techniques are likely to be useful are mentioned.

32. Most statistical offices keep some statistics about domestic migration patterns. If these migration events can be matched to some other data about the person or household that has migrated, an analysis of which categories of persons are likely to migrate (to cities, suburbs, country side) could be undertaken. Some sort of budget survey (i.e. about consumers' expenditure patterns) is available in most statistical offices; it could go beyond the collection, categorization and tabulation of the material, and try to assess the major determinants of expenditure preferences and shifts between the main competing alternatige destinations.

33. From a business register, one could try to find out what are the major factors which determine whether astarting business survives the first year. Data from the Employment Offices, when matched with additional information about the job seekers, could be used to find out the different unemployment patterns over time in several groups of people.

[13] Data are called structured when some variables show a systematic non-noisy relation.

References

AAR97 E. Aarts, J.K. Lenstra (eds.): Local search in Combinatorial Optimization. John Wiley, 1997.

ACM96 a special issue with several articles on Data Mining: *Communications of the ACM*, nov. 1996, vol 39, no 11.

CLE94 W.S. Cleveland: The elements of Graphing Data (Revised ed.). Hobart Press, 1994.

FAY96 U.M. Fayad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.): Advances in Knowledge Discovery and Data Mining. MIT Press, 1996.

HEC97 D. Heckerman, D. Pregibon, R. Uthurusamy (eds.): Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (Newport Beach, CA, Aug. 14-17,1997), AAAI Press, 1997.

ISR82 Israëls A.Z., Bethlehem J.G., Driel J. van, Jansen M.E., Pannekoek J., Ree S.J.M. de, Sikkel D.: Multivariate analysis methods for discrete variables. *Metron*, 40, 1982, pp. 193-212.

MIL81 R.G. Miller: Simultaneous Statistical Inference (2nd ed.) Springer, 1981.

TUK77 J.W. Tukey: Exploratory Data Analysis. Addison Wesley, 1977.