MULTIDIMENSIONAL DETECTIVE

Alfred Inselberg,* Multidimensional Graphs Ltd†
&
Computer Science Department
Tel Aviv University, Israel
aiisreal@math.tau.ac.il

Abstract

The display of multivariate datasets in parallel coordinates, transforms the search for *relations* among the variables into a 2-D pattern recognition problem. This is the basis for the application to *Visual Data Mining*. The Knowledge Discovery process together with some general guidelines are illustrated on a dataset from the production of a VLSI chip. The special strength of parallel coordinates is in modeling **relations**. As an example, a simplified Economic Model is constructed with data from various economic sectors of a real country. The visual model shows the interelationship and dependencies between the sectors, circumstances where there is competition for the same resource, and feasible economic policies. Interactively, the model can be used to do trade-off analyses, discover sensitivities, do approximate optimization, monitor (as in a Process) and Decision Support.

Introduction

In Geometry parallelism, which does not require a notion of angle, rather than orthogonality is the more fundamental concept. This, together with the fact that orthogonality "uses-up" the plane very fast, was the inspiration in 1959 for "Parallel" Coordinates. The systematic development began in 1977 [4]. The goals of the program were and still are (see [6] and [5] for short reviews) the visualization of multivariate/multidimensional problems without loss of information and having the properties:

1. Low representational complexity. Since the number of axes, N equals the number of dimensions (variables) the complexity is $O(N)$,
2. Works for any N,
3. Every variable is treated uniformly (unlike "Chernoff Faces" and various types of "glyphs"),
4. The displayed object can be recognized under projective transformations (i.e. rotation, translation, scaling, perspective),
5. The display easily/intuitively conveys information on the properties of the N-dimensional object it represents,
6. The methodology is based on rigorous mathematical and algorithmic results.

Parallel coordinates (abbr.||-coords) transform multivariate relations into 2-D patterns, a property that is well suited for Visual Data Mining.

*Senior Fellow San Diego SuperComputing Center
†36A Yehuda Halevy Street, Raanana 43556, Israel

0-8186-8189-6/97 $10.00 © 1997 IEEE

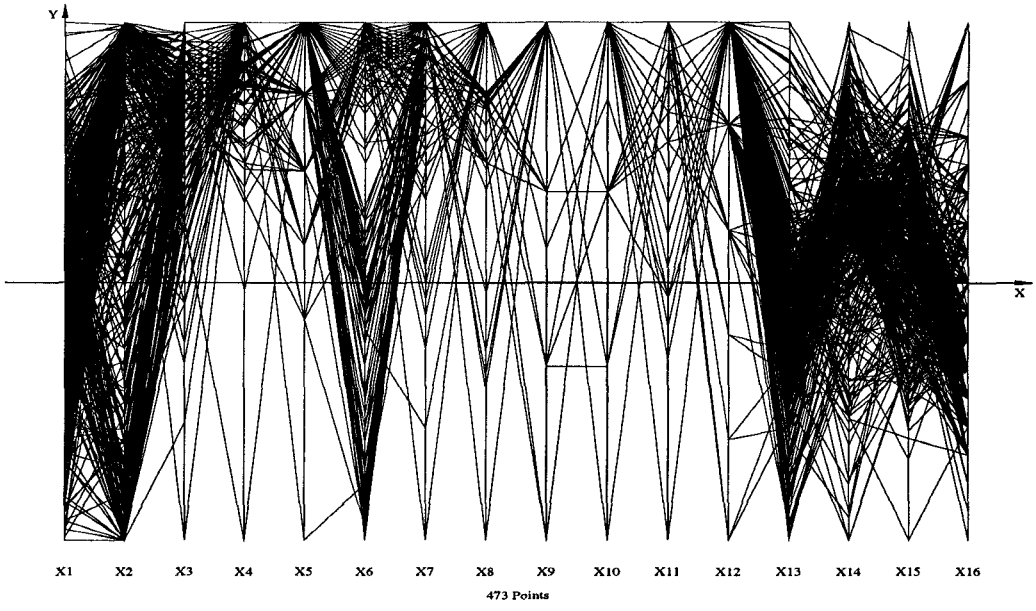

Figure 1: The full dataset consisting of 473 batches

Several Data Mining tools EDA (Chomut [2]), Finsterwalder[3], VisuLab(Hinterberger[10]), ExplorN(Carr et al), Influence Explorer(Spence & Tweedie [11]), WinViZ(Eickemeyer), VisDB(Keim [7]), Xmdv(Ward[8]), XGobi(Buja), Strata(||-coords by Gleason), Diamond(Rogowitz et al [9]), PVE(Inselberg, Adams, Hurwitz, Chatterjee, Austel) etc. include ||-coords. Here we focus on the Data Mining application and describe:

- A scenario for the discovery process,
- Guidelines for using ||-coords in Data Mining, and
- The construction and use of visual models for multivariate relations.

There are certain basics (see references) which have important ramifications. For example, due to the Point $\longleftrightarrow$ Line *duality*, some actions are best performed in the dual and their opposite in the original representation. Another important matter is the design of queries. The task is akin to accurately cutting complicated portions of an N-dimensional "watermelon" (i.e. the N-dimensional representation of the dataset). The "cutting tools" are the queries which must also operate in the dual (i.e. the ||-coords display). They need to be few, exquisitely well chosen and intuitive. This requires an efficient and convenient way of combining the *"atomic"* queries to form complex queries, corresponding to more intricate "cuts" of the dataset; and there are other issues. These points are not often appreciated and, as a result, software usually mimic the experience derived from the standard and more familiar displays (i.e. not the dual), rather than exploit the special strengths of the methodology and avoid its weaknesses.

Without the proper **geometrical** understanding and queries, the effective use of ||-coords becomes limited to small datasets. By contrast, skillful application of the methodology's strengths enables the analysis of datasets consisting of thousands of points and hundreds of variables. The intent here, is not to elaborate on the design and implementation criteria but rather to provide some insights on the "discovery process". The paradigm is that of a detective, and since many parameters (equivalently dimensions) are involved we really mean a "multidimensional detective".

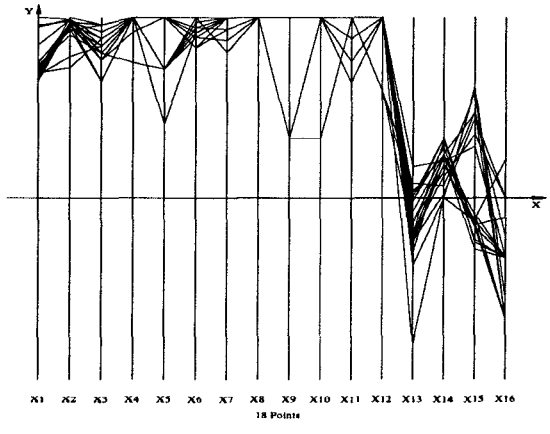

Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

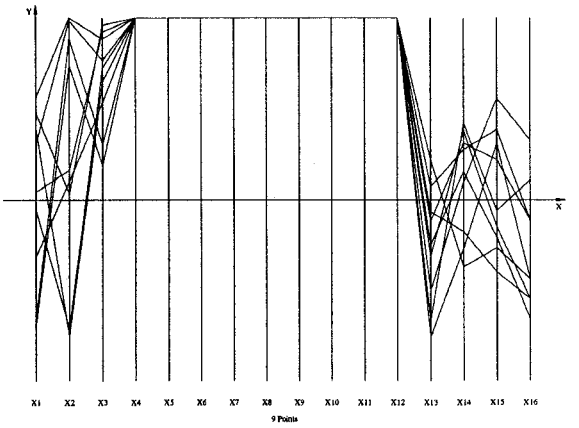

Figure 3: The batches with zero in 9 out of the ten defect types.

The Problem

Aside from starting the exploration without bias, together with some healthy scepticism about the "convictions" of the domain experts, the first admonition is:

- **do not let the picture intimidate you,**

as can easily happen by taking an uninformed look at Fig. 1 where our subject's dataset is displayed. It pertains to the production data of 473 batches of a VLSI chip with measurements of 16 process parameters denoted by $X1, X2, \ldots, X16$. The *yield*, as the % of useful chips produced in the batch, is denoted by $X1$, and $X2$ is a measure of the *quality* (in terms of speed performance). Ten different types of *defects* are monitored and

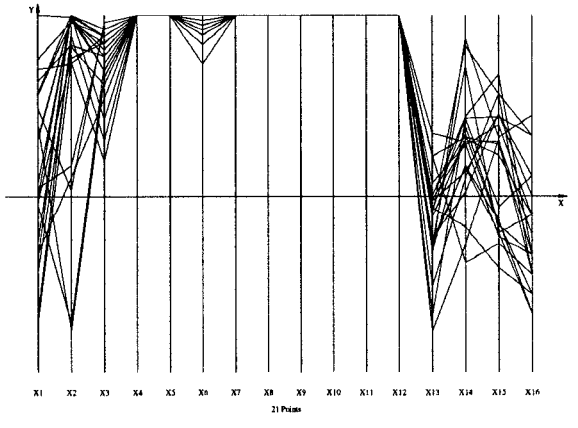

Figure 4: The batches with zero in 8 out of the ten defect types.

the variables' scales of $X3$ through $X12$ are inverted so that 0 (zero) amount appears at the top. The remaining, $X13$ through $X16$, denote some physical parameters. We emphasize that this is a *real* dataset and in order to protect the innocent, as well as confuse the competition, it is not possible to give a more explicit description or show numerical values; for our purposes it is also not necessary. Prior to embarking on our exploration it is essential to

- **understand the objectives and use them to obtain "visual cues".**

Here the objective is to raise the yield, $X1$, and maintain high quality, $X2$, a multiobjective optimization (since more than one objective is involved). Production experts believed that it was the presence of defects which hindered high yields and qualities. So the goal was to achieve *zero defects*.

Discovery Process - How to be a Multidimensional Detective

The keen observer can ascertain from Fig. 1 the distributions, with $X2$ being somewhat bipolar (having higher concentrations at the extremes), and $X1$ having something like a normal distribution about it's median value. This brings us to the next admonition. Namely, no matter how messy it looks,

- **carefully scrutinize the picture**

and you are likely to find some patterns, let's call them *visual cues*, which hint at the relations among the variables.

We embark on our quest and the result of our first query is shown in Fig. 2 where the batches having the highest $X1$ and $X2$ have been isolated. This in an attempt to obtain *clues*; and two real good ones came forth (the visual cues we spoke of). Notice $X15$ where there is a separation into two clusters. As it turns out, this gap yielded important (and undiscloseable) insight into the physics of the problem.

The other clue is almost hidden. A careful comparison – and here interactivity of the software is essential – between Fig. 1 and Fig. 2 shows that some batches which were high in $X3$ (i.e. and due to the inverted scale low in that defect) *were not included in the selected subset.* That casts doubt

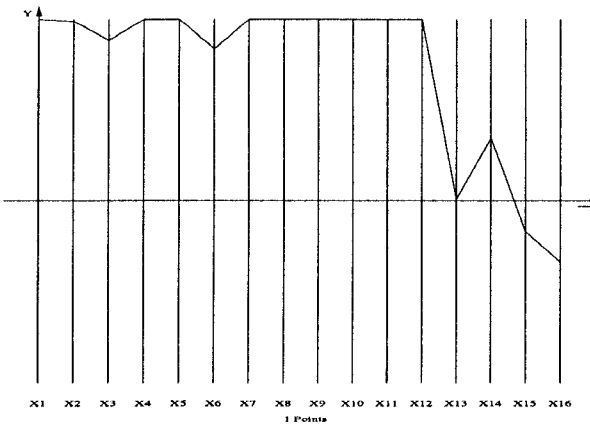

Figure 5: The best batch. Highest in Yield, $X1$, and very high in Quality, $X2$.

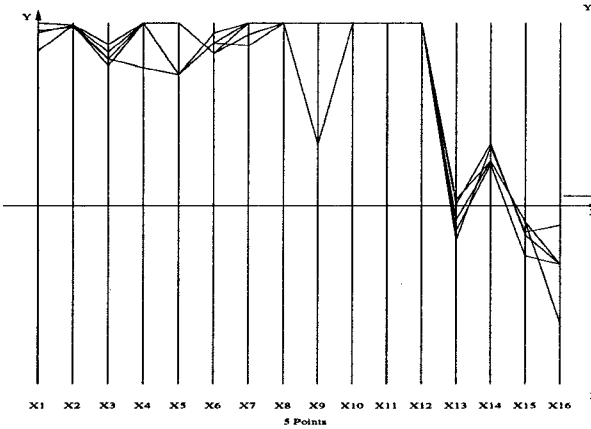

Figure 6: Batches with the highest Yields do not have the lowest defects in $X3$ and $X6$.

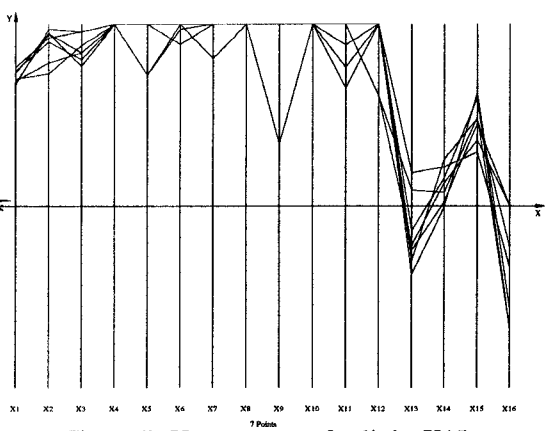

Figure 7: Upper range of split in $X15$

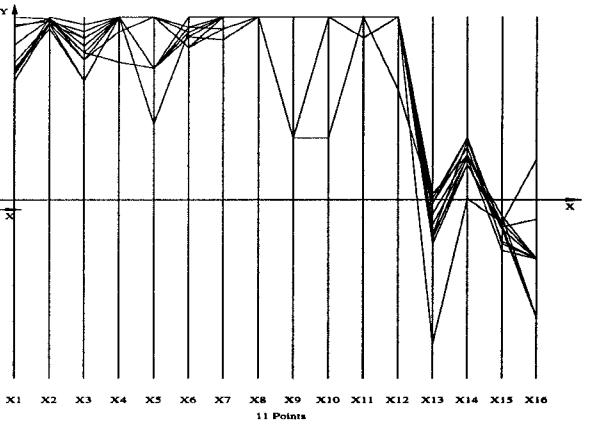

Figure 8: Batches with the lower range of $X15$

into the belief that zero defects are the panacea, and motivates the next query where we search for batches having zero defects in at least 9 (excluding $X3$ where we saw that there are problems) out of the 10 categories. The result is shown in Fig. 3 and is a shocker. There are 9 such batches and *all of them have poor yields and for the most part also low quality*! What is one to do? Refer again to the previous admonition and scrutinize the original picture Fig. 1 for visual cues relevant to our objectives and our findings so far. And ... there is one staring us in the face – the visual difference between $X6$ and the other defects. It shows that the process is much more sensitive to variations in $X6$ than the other defects. We chose to treat $X6$ differently and remove its zero defect constraint. The result is seen in Fig. 4 and, remarkably, the very best batch (i.e. highest yield with very high quality) is included. This is an opportunity to learn, so when that batch is highlighted with a different color (can't be seen in black and white) *it does not have zeros (or the lowest values) for* $X3$ *and* $X6$ as shown separately in Fig. 5. A "heretical" finding, but perhaps this is due to measurement errors in that one data item. We return to the full dataset and isolate the cluster of batches with the top yields (note the gap in $X1$ between them and the remaining batches). These are shown in Fig. 6 and they confirm that small amounts (the ranges can be clearly delimited) of $X3$ and $X6$ type defects are *essential* for high yields and quality. The moral of the story is

- **test the assumptions and especially the "I am really sure of ..."s.**

Let us return to the subset of data which best satisfied the objectives, Fig. 2, to explore the gap in the range of $X15$. In Fig. 7 we see that the cluster with the high range of $X15$ gives the lowest (of the high) yields $X1$, and worse it does not give *consistently* high quality $X2$. Whereas the cluster corresponding to the lower range, Fig. 8, has the higher qualities and the full range of the high

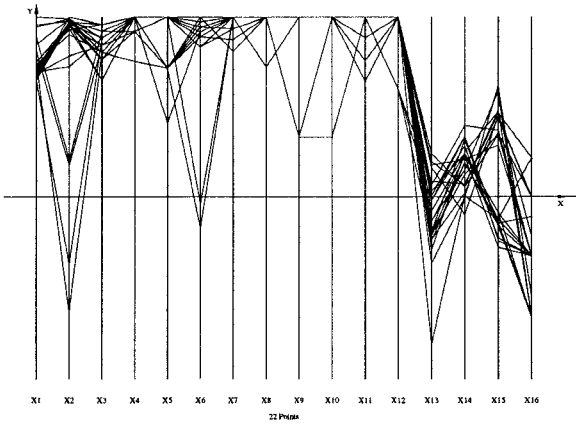

Figure 9: Top Yields produce split in $X15$

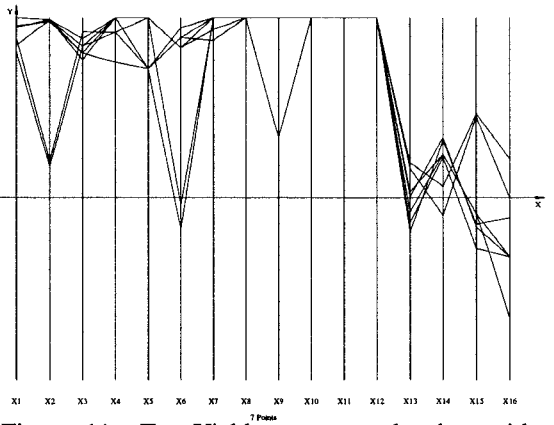

Figure 11: Top Yields, note two batches with lower Quality

yield. It is evident that the small ranges of $X3$, $X6$ close to (but not equal to) zero, together with the short (lower) range of $X15$ provide *necessary* conditions for obtaining high yields and quality. Using a characterization algorithm it can be shown that these conditions are also *sufficient*. Given any subset of the data the algorithm finds:

1. the smallest subset of variables which describe the data without loss of information, and
2. orders the variables in terms of their predictive power.

In our case these are the 3 parameters $X3$, $X6$ and $X15$ which are needed for the characterization of a very good batch. By a stroke of good luck these 3 can be checked *early* in the process avoiding the need of "throwing good money after bad". Looking again at Fig. 1 we notice a gap in $X1$ between the top 5 batches and the rest. The high cluster consists of *only* those having the small amounts of $X3$ and $X6$ and lower range of $X15$. This bit of serendipity provides an instance of my favorite "stochastic" theorem :

- **you can't be unlucky all the time!**

Some Deeper Insights

Why the gap in $X15$? It was obtained by imposing *simultaneously* the constraints for top yields and quality. Fig. 9 shows the result of constraining only $X1$ and the resulting gap in $X15$, whereas the high $X2$ by itself does not yield a gap as shown in Fig. 10. And this, I am told, provided further insights into the physics of the problem.

Just for fun we look next into the very top yields

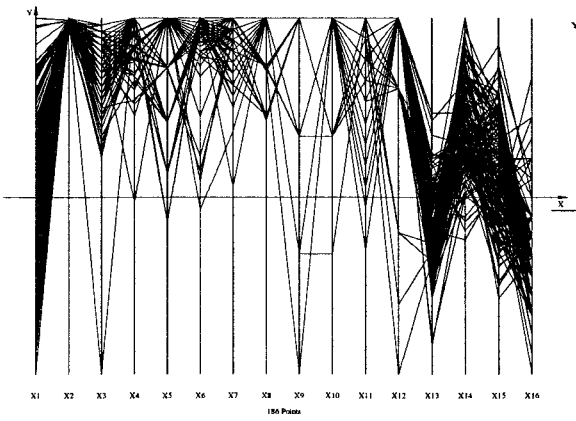

Figure 10: High $X2$ does not cause split in $X15$

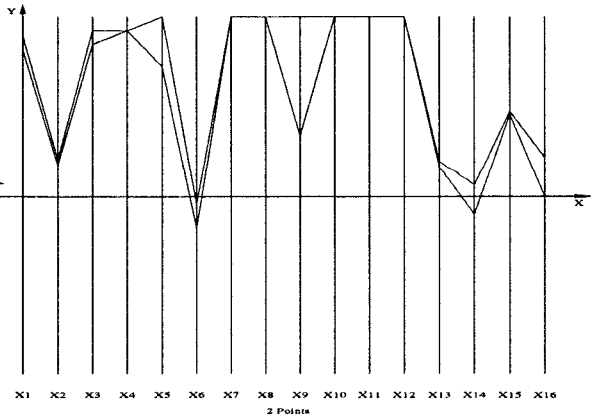

Figure 12: The two batches with high $X1$ and lower $X2$

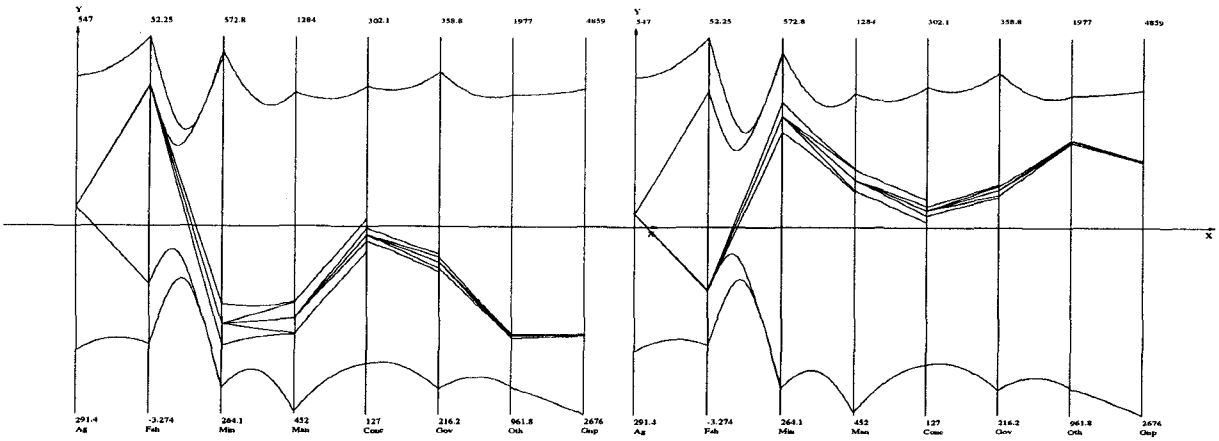

Figure 13: Model of a country's economy

Figure 14: Competition for labor between the Fishing & Mining sectors – compare with previous figure

Fig. 11 and see that except for two batches the others also have very high $X2$. Isolating the lower quality batches turns out to be very informative. The picture, Fig. 12, suggests that high yields and lower quality may be due to the different ranges of $X6$, whose influence we have already seen elsewhere, and specific ranges of $X13$, $X14$, $X15$, $X16$. This observation suggests that it may be possible to partition this multivariate problem into sub-problems pertaining to the individual objectives. The wide variation of $X9$ for these two batches, as seen in Fig. 12, lead to further testing and the conclusion that $X9$ is a "junk variable" with respect to the objectives. The observations and conclusions here involved a relatively small number of variables and should be cross-checked with larger datasets. In general, a dataset with P points has 2^P subsets anyone one of which can be the "interesting" one (with respect to the objectives). Our approach can provide a powerful tool for coping with this combinatorial explosion. The **visual cues** obtained can help to rapidly focus on the interesting portions of the data. In this sense, it can serve as a **preprocessor** to other methods, in addition to providing unique insights on its own.

After this analysis, it was revealed that this was a well studied problem and dataset, and our findings differed markedly from those found with other methods for process control [1]. This is not an isolated case and there have been other successful application of this approach in the manufacture of printed card boards, PVC and manganese production, retailing, finance, trading, insurance, seasonal weather forecasts, risk analysis, determination of skill profiles" (i.e. as in drivers, pilots etc) and elsewhere. The results frequently surprised the domain experts.

Visual & Computational Models

We have outlined a process for discovering interesting, with respect to the objective, **relations** among the variables in multivariate datasets. The real strength of the methodology is the ability to construct and display such relations in terms of hypersurfaces – just as we model a relation between two variables by a planar region. Then by using an interior point algorithm with the model we can do trade-off analyses, discover sensitivities, understand the impact of constraints, and in some cases do optimization. We just want to indicate how this works. For this purpose we use a dataset consisting of the outputs of various economic sectors and other expenditures of a particular (and real) country. It consists of the monetary values over several years of the **Agricultural** output, outputs of the **Fishing**, **Mining**, **Manufacturing** and **Construction** industries, together with **Government**, Miscellaneous spending and resulting GNP; eight variables altogether.

We will not take up the full ramifications of constructing a model from data. Rather, we want to illustrate how ||-coords may be used as a **visual** modeling tool. Using a Least Squares technique we "fit" a function to this dataset; for our purposes here we are not concerned whether the choice of function is "good" or not. The specific function we obtained bounds a region in R^8 and is represented by the upper and lower curves(envelopes) shown in Fig. 13(the interested reader may want to refer to the previously cited references).

The picture is in effect a visual model of the

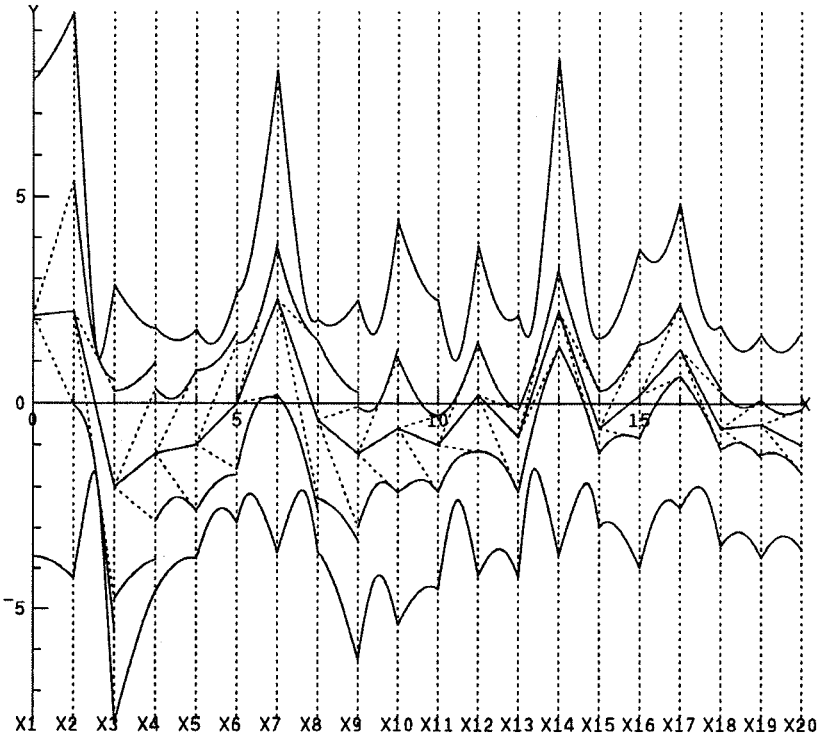

Figure 15: A Convex Hypersurface in 20-D and Interior Point Algorithm.

country's economy, incorporating it's capabilities and limitations, interelationships among the sectors etc. A point interior to the region, satisfies all the constraints simultaneously, and therefore represents (i.e. the 8-tuple of values) a *feasible economic policy* for that country. Using an interior point algorithm (see previously cited references) we can construct such points. It is done interactively by sequentially choosing values of the variables and we see the result of one such choice in Fig. 13. Once a value of the first variable is chosen (in this case the agricultural output) within it's range, the dimensionality of the region is reduced by one. The upper and lower curves between the 2nd and 3rd axes correspond to the resulting 7-dimensional hypersurface and show the *available* range of the second variable (Fishing) reduced by the constraint (i.e. fixing the value of the first variable). In fact, this can be seen (but not shown here) for the rest of the variables. That is, due to the relationship between the 8 variables, a constraint on one of them impacts all the remaining ones and restricts their range. The display allows us to experiment and actually see the impact of such decisions "downstream". By interactively varying the chosen value for the first variable we found, from this model, that low values for Agriculture correspond to low ranges of values for Fishing, and similarly corresponding to high values occurring together. So it is not possible to have a policy that favors Agriculture without also favoring Fishing and vice versa. The algorithm fails where at any stage the polygonal line crosses an intermediate curve, and that is very informative.

Proceeding, a very high value from the available range of Fishing is chosen next. It corresponds to very low values of the Mining sector. By contrast in Fig. 13 we see that a low value in Fishing yields high values for the Mining sector. This inverse relation was investigated and it was found that the country in question has a large number of migrant workers. When the fishing industry is doing well most of them are attracted to it leaving few available to work in the mines and vice versa. The comparison between the two figures shows the *competition for the same resource* between Mining and Fishing. It is especially instructive to discover this interactively. The construction of the interior point proceeds in the same way.

Let us move over to Fig. 15 where the same construction is shown but for a more complex 20-dimensional hypersurface ("model"). The intermediate curves (upper and lower) also provide

valuable information and "previews of coming attractions". They indicate a neighborhood of the point (represented by the polygonal line) and provide a feel for the local curvature. Note the narrow strips between $X13$, $X14$ and $X15$ (as compared to the surrounding ones), indicating that for this choice of values these 3 are the *critical* variables where the point is "bumping the boundary". A theorem guarantees that a polygonal line which is in-between all the intermediate curves/envelopes represents an interior point of the hypersurface and all interior points can be found in this way. If the polygonal line is tangent to anyone of the intermediate curves then it represents a *boundary point*, while if it crosses anyone of the intermediate curves it represents an *exterior point*. The latter enables us to see, in an application, the first variable for which the construction failed and what is needed to make corrections. By varying the choice of value over the available range of the variable interactively, sensitive regions (where small changes produce large changes downstream) and other properties of the model can be easily discovered. Once the construction of a point is completed it is possible to vary the values of each variable and see how this effects the remaining variables. So one can do *trade-off analysis* in this way and provide a powerful tool for, Decision Support, Process Control and other applications.

It should be self-evident that the efficacy of a visual data mining tool needs to be judged by applying it to **real** and necessarily challenging datasets. Flashy demos based on artificial or small datasets can be very impressive but misleading. Each multivariate dataset and problem has its own "personality" requiring substantial variations in the discovery scenarios and calls for considerable ingenuity – a characteristic of good detectives. It is not surprising then that the most frequent requests are for tools to, at least partially, automate the exploration process. Such a development is under way and will include a number of new features, including **intelligent agents**, gleaned from the accumulated experience.

References

[1] E.W. Bassett. Ibm's ibm fix. *Industrial Computing*, 14(41):23–25, 1995.

[2] T. Chomut. *Exploratory Data Analysis in Parallel Coordinates*. M.Sc. Thesis, UCLA Comp. Sc. Dept., 1987.

[3] R. Finsterwalder. *A Parallel Coordinate Editor as a Visual Decision Aid in Multi-Objective Concurrent Control Engineering Environment 119-122*. IFAC CAD Contr. Sys., Swansea, UK, 1991.

[4] A. Inselberg. *N-Dimensional Graphics, Part I – Lines and Hyperplanes, in IBM LASC Tech. Rep. G320-2711, 140 pages*. IBM LA Scientific Center, 1981.

[5] A. Inselberg. *Parallel Coordinates : A Guide for the Perplexed, in Hot Topics Proc. of IEEE Conf. on Visualization, 35-38*. IEEE Comp. Soc., Los Alamitos, CA, 1996.

[6] A. Inselberg and B. Dimsdale. *Parallel Coordinates: A Tool For Visualizing Multidimensional Geometry, in Proc. of IEEE Conf. on Vis. '90, 361-378*. IEEE Comp. Soc., Los Alamitos, CA, 1990.

[7] D. A. Keim and H. P. Kriegel. Visualization techniques for mining large databases: A comparison. *Trans. Knowl. and Data Engr.*, 8-6:923–938, 1996.

[8] Ward M. O. *XmdvTool: integrating multiple methods for visualizing multivariate data, Proc. IEEE Conf. on Visualization, San Jose, CA, 326-333*. IEEE Comp. Soc., Los Alamitos, CA, 1994.

[9] M. Schall. Diamond and ice : Visual exploratory data analysis tools. *Perspective, J. of OAC at UCLA*, 18(2):15–24, 1994.

[10] C. Schmid and H. Hinterberger. *Comparative Multivariate Visualization Across Conceptually Different Graphic Displays, in Proc. of 7th SSDBM*. IEEE Comp. Soc., Los Alamitos, CA, 1994.

[11] L.A. Tweedie, R. Spence, H. Dawkes, and Su H. *Externalizing Abstract Mathematical Models, Proc. CHI, Vancouver, Can., 406-412*. ACM Press, 1996.