An Application of the Empathic Visualisation Algorithm (EVA)

A. Loizides and M. Slater

Department of Computer Science,
UCL University of London,
Gower Street,
London WC1E 6BT, UK.

Abstract

Vision is the sense with the largest bandwidth. This paper demonstrates EVA, a technique that would allow perception to amplify cognition in the visualisation of multidimensional data sets. Visualising all the interrelationships of the variables in a data set holistically is a challenging task. We apply Genetic Programming techniques to this problem to map quantitative measurements over the variables of the data set, to qualitative measurements of the visual structure that are significant to human perception. The procedure describes a generic approach to map multidimensional data sets to visual structures. Initial results indicate that the method performs significantly better than a random mapping technique and shows great promise.

Key Words: Information Visualisation, multivariate data, multidimensional visualisation, naturalistic representations, data visualisation, genetic programming.

1 Introduction

This paper describes research that is investigating the use of naturalistic visual structures for representation of multivariate data sets with the aim of increasing insight about our data: 'the purpose of visualisation is insight, not pictures'[3]. The data we are interested at has the following attributes: the data is large, multidimensional, of non-physical nature and has hidden information. Examples include, financial data, business information, collections of documents and abstract conceptions. Such information has no obvious spatial mapping.

Instead of processing individual details and have as our output numbers and text, we examine the use of the visual system to process visual structures holistically and thus obtain an overall global view of a data set. We are gaining information out of an overall view of qualitative measurements. It allows us to examine the important features of the visual structures very quickly, and also notice abnormalities very quickly. It can be viewed as the process of "using vision to think", or the "mind's eye" [3] to gain insight into what are often complex abstract data systems.

The method is based on a technique derived by Slater [12] to automatically map data to visual structures using genetic programming techniques. The objective of EVA is, given a data set and an observer, to construct a visualisation such that salient features of the data can be intuitively recognised by the observer. In other words, humans can detect patterns that reveal the underlying structure in the data more readily than a direct analysis of the numbers would. In order to achieve this we use visual structures that are *naturalistic* in the sense that no human requires special knowledge for interpretation and also the mapping from data to the visual structure is *automated*. The mapping should be such that the 'important' features of the data set are mapped to features of the visual structure that are 'important' and significant to human perception or human emotions.

EVA, the method introduced here to visualise complex data sets, can be thought of as the initial step to establish a clearer pictorial representation of a problem. It is by no means an alternative to statistical analysis, but a complementary one. It can be seen as the front line of attack revealing intricate structure in the data.

Section 2 is devoted to a brief description of the related work. In section 3 the method is explained in detail, followed by initial results that proof the method correct in section 4. Conclusions and Future Work is covered in section 5.

2 Related Work

Interesting visualisations start with multivariate data sets that have too many variables to be encoded directly using 1d, 2d and 3d visual structures. For this kind of data, graphs and charts lose their effectiveness. More accurately we can say that for data sets with more than 3 variables an orthogonal visual structure is not sufficient.

Most of the existing techniques to visualise multivariate data sets use abstract visual structures. For example colour encoded visualisations, sophisticated graph-based visualisations or glyph-based visualisations. These have the following drawbacks. It is very hard, if not impossible, to get an overall view of your data set. It requires learning time from the point of view of a user, and are slow if used for decision making because of the wide range of functionality that needs to be performed in order to find more about the data set. Most of the functionality is included because users cannot get a holistic view of their data set in one picture. Some visualisation methods make use of a naturalistic visualisation but the mapping performed from data to the visual structure is user specified and arbitrary. It does not take into account the impact of the visual structure on the emotions of the observer.

Mihalisin, Timlin and Schwegler [10] developed a technique to visualise a scalar dependent variable that is a function of many 'independent' variables. Each variable is plotted within the space delimited by the previous variable for each discrete value that represents the variables range. The process is done recursively until the whole function has been plotted. It is a hierarchical technique to visualise and visually analyse multivariate functions, data and distributions in various ways. The order you place your variables in the hierarchy affects the visualisation you get. Some are comparatively easier to understand than others. Therefore it gets quite complicated as the dimensionality of the variables increases and as the range of values of your variables increases.

Bertin [2] developed a direct technique for creating multidimensional visual structures from multivariate data sets called 'permutation matrices'. It involves representing rows of data as bar charts and sorting them. Graphical depictions were placed on cards and permuted with metal rods with the aim to produce patterns, therefore, clustering similar cases.

Beshers and Feiner [5] described a technique called 'worlds within worlds', a multidimensional visual structure based on overloading. They visualise high-dimensional functions by placing 3d coordinate systems inside other 3d coordinate systems recursively until all dimensions are included. Changing the position of the inner coordinate system results in changes in the surface displayed since three variables are changed. However, at any one time, the surface displayed is constructed out of 3 variables (the outer coordinate system) and the constant values of the rest of the variables.

Inselberg [6] proposed a method he refers to as 'parallel coordinates'. The fact that orthogonality 'uses-up' the plane very fast and also that parallelism does not require a notion of angle led to the inspiration of parallel coordinates. It can be seen as the method to model relations, or a 2d pattern recognition problem. Parallel coordinates transform multivariate relations into 2d patterns. However, the method requires skill from the user to get a geometrical understanding and properly query the picture.

Spoerri [13] built a system called InfoCrystal as a tool for information retrieval. It can be used as a visualisation tool and a visual query language. It provides facilities to visualise all the possible discrete as well as continuous relationships among certain amount of concepts using a generalisation of the Venn diagram that we are all familiar with. For the discrete case, InfoCrystal uses attributes like proximity, rank, shape colour and size for encoding and for the continuous case a polar representation used with a qualitative sense of how ranked documents are related to the input criteria.

Naturalistic graphical representations has not been used that much mainly because of the subjectiveness of the visual structure. However, Alexa and Muller, [1] make use of more naturalistic representations in a technique called Visualisation by Examples. Based on the specification of only a few correspondences between data values and visual representations, complex visualisations are produced. The basis of their method is their 'morphing' technique (for more about morphing see Alexa and Muller [1]). Morphing between two graphical objects results in a one dimensional space. Morphing between an element of such space and a third graphical base object results in a two dimensional space. By repeating this process they construct a space of any given dimension.

Chernoff, [4] proposed a well known technique that uses such naturalistic visual scales. He used simple schematic faces for the visualisation of multivariate data, by assigning variates in the data set to features of the face. However, not every variable is treated uniformly in this approach.

The use of faces as a mean for representing multivariate data sets was investigated further. Wilkinson, [14] has shown that when considering similarity of data sets, faces prove to be better representation than other forms. Walker, [7] also explained that the use of faces as a means of communication leads to fewer mistakes being made and to users becoming more involved in what they are doing. Moreover, Walker has shown that the use of a human face does not introduce an extra element that could distract a user from the task they are trying to achieve, thereby potentially reducing their effectiveness. Instead, it actually increases the impact of the information, while providing a familiar medium for communication. Furthermore, the use of cartoon faces has been proved successful in improving the communication of financial and accounting information, [11] [9].

Most of the techniques described above do not use naturalistic graphical representations [10],[2], [5], [6], [13] but rather an abstract representation. The other techniques [1],[4] although they use a naturalistic representation, a face, the mapping from data to the visual structure is an arbitrary one.

EVA, the method applied here explores the use of naturalistic representations and also describes an automatic mapping from data to visual structure that takes into account the impact of the visual structure on the emotions of the observer. A fundamental goal of this mapping is that the perceptually or emotionally significant features of the visual structure directly reflect the *value system* that reflects the interest or importance or consequences of aspects of the data for the situation of the observer. We call this goal *visual homomorphism*. Hence, the mapping is constructed in such a way that this visual homomorphism is realised.

3 Method

This section describes our approach used to visualise multivariate data. Before explaining how we automatically map data streams to visual structures using genetic programming techniques some notation and assumptions will be stated. For a more detailed explanation of the method see Slater [12].

First of all, consider the representation of multivariate data in an $n \times k$ data matrix X, consisting of n observations and k variables $x_1, x_2, ..., x_k$. Each row in the data matrix typically represents an individual and there are k observations made for each individual.

Let $\nu_s(X)$, $s = 1, 2, ..., p$ be p functions over the data representing 'values' over the data matrix.

In addition, consider a visual structure (Ω). Similarly there are p importance aspects of the Ω that are measurable and significant to human perception or emotions, $e_s(\Omega)$, $s = 1, 2, ..., p$.

The fundamental goal, in terms of any X and any Ω, is to produce a mapping $\mu(X) \longrightarrow \Omega$ such that 'values' over the data matrix are reflected in characteristics of the visual structure. A 'characteristic' is a measurement of some aspect of Ω as a whole (such as the emotions on a face) rather than some individual feature (such as the shape of the mouth).

Moreover, suppose that there are r features of the visual structure: $\phi_t(\Omega)$, $t = 1, 2, ..., r$. Knowing these features of Ω results in being able to render it. Once rendered, we can measure it to determine its characteristics $e_s(\Omega)$, $s = 1, 2, ..., p$.

Finally, we introduce *feature functions* over the data matrix: $f_t(X)$, $t = 1, 2, ..., r$. These functions completely determine the features of the visual structure, in fact

$$f_t(X) = \phi_t(\Omega),\ t = 1, 2, ..., r \tag{1}$$

The values of these functions are *interpreted* as the values of the features of the visual structure. Our goal is to choose these functions f in order to attain the required correspondance between the value system over the data and the characteristics of the visual structure and therefore, to attain the visual homomorphism.

Let $\nu = (\nu_1, \nu_2, ..., \nu_p)$ and $e = (e_1, e_2, ..., e_p)$. Suppose $\| \nu - e\|$ is a measure of the 'distance' between these two vectors. Then the specific goal is to choose $f_t(X)$, $t = 1, 2, ..., r$ such that $\| \nu - e\|$ is minimised.

3.1 Using Genetic Programming (GP)

The minimisation problem introduced above can be tackled with a GP. Let $F_i = (f_{1i}, f_{2i}, ..., f_{ri})$ be a specific set of feature functions which when applied to X produce the visual structure features. Choose at random a large collection of such sets of functions F_i, $i = 1, 2..., N$. This defines a *population* of sets of feature functions. The ith member of the population produces a specific visual structure Ω_i. This visual structure has characteristics $e_s(\Omega_i)$, $s = 1, 2, ..., p$. These characteristics can be used to produce the distance measurement $\| \nu - e\|$. The distance measurement can be used to compute a 'fitness' for the ith member of the population. Hence, each member of the population has an associated fitness, which can be expressed as a probability. These probabilities determine survival into the next generation and selection for mating - thus producing a second generation. The process continues until (possible) convergence. The most fit member of all populations is chosen for the required mapping.

3.2 Overview

A summary of the steps required in order to implement the method is described below, using the notation above.

1. Decide on Ω, the visual structure. Determine the number of features Ω has. Assuming there are r you need $f_t(X)$, $t = 1, 2..., r$ feature functions in order to render an individual.

2. The observer identifies the p important values of the data set $\nu_s(X)$, $s = 1, 2, ..., p$.

3. Identify $e_s(\Omega)$, $s = 1, 2, ..., p$ the p characteristics of the visual structure that measure its totality and are significant to human emotions and perception.

4. Identify the fitness function i.e. the function we are trying to minimise, which is, for example:

$$\sum_{d=1}^{n} \sum_{s=1}^{p} (\nu_{ds} - e_{ds})^2 \tag{2}$$

5. Identify the GP parameters and run the GP.

The steps shown above achieve the visual homomorphism described earlier. In other words, it can be defined as the 'extraction and visualisation of qualitative data from quantitative one'. We will illustrate the procedure used, by the means of an example using a set of three circles as the visual structure. This has been chosen as a simple example, but where the outcome is intuitively obvious. A set of circles as in figure 1 is more ordered than the set of circles in figure 2. It is implemented to test the effectiveness of this approach. Results from a user experiment using the method will also be given in a following section.

3.3 Example

First of all, we need to obtain the data set we want to visualise. Consider a simulated multivariate financial data set. Table 1 shows an example of such data.

We will now describe the steps required to run the GP that produced the automatic mapping from data to visual structure.

3.3.1 Step 1

For the particular example, there are 12 feature functions $f_t(X)$, $t = 1, 2, .., 12$. This 12 functions will correspond to the 12 features $\phi_t(\Omega)$, $t = 1, 2, ..., 12$ that determine Ω.

Company	No. of shares	Earnings	Dividend payout	Liquid Assets	Liabilities	Costs	Sales	Staff
1	0.12	0.25	0.70	0.60	0.61	0.21	0.20	0.11
2	0.34	0.15	0.24	0.82	0.44	0.74	0.31	0.05
.	.	.	.	.	.	.	.	.
N	.	.	.	.	.	.	.	.

Table 1: Example of the self-created data set.

- Six values for the three pairs of x, y coordinates representing the centre of the circles.
- Three values representing the sizes of the three circles.
- Three values representing the grey scale value, from 0 to 1, for the colour of the circles.

3.3.2 Step 2

The three measurements over the data that are of interest to us are the following $\nu_s(X)$, $s = 1, 2, 3$:

$$Profit = \frac{earnings}{costs} = X_1 \tag{3}$$

$$Dividend\% = \frac{dividendpayout}{earnings - costs} * 100 = X_2 \tag{4}$$

$$Liquidity = \frac{liquidassets}{liabilities} = X_3 \tag{5}$$

We also need a scoring system for these measurements. Each of the companies will be scored on a scale of 0 to 100 under each category. The score is based on what financial experts consider a reasonable value in each case:

Profit(ν_1) It bounds from $0 - 100$, so that when earning are five times the cost then a score of 100 is given, if it is equal or less the cost, then a score of 0 is given.

Dividend(ν_2) %: It bounds from $0 - 100$, so that when the ratio is 30% or higher then a score of 100 is given, if it is 0% then a 0 is given.

Liquidity(ν_3) It also bounds from $0 - 100$, so that when the ratio of liquid assets to liabilities is 1.5 or more then a score of 100 is given, if 1 or below then a score of 0 is given.

3.3.3 Step 3

We also need to identify the aspects of the visual structure that are measurable and significant to human perception $e_s(\Omega)$, $s = 1, 2, 3$. For this particular example, the following aspects in order of importance will be considered.

- The distance between the centres of the circles - mapped to profit. The closer to concentric the circles are the more profit is indicated.
- The proportionality of the radii - mapped to dividend %. The closer to an arithmetic progression the circles are the better the dividend % is.
- The proportionality of the colour - mapped to liquidity. The closer to an arithmetic progression the colours of the circles are, independently to which circle they belong, the better the company is performing in terms of profitability.

As with the data set, we also need a scoring system.

1. e_1: The maximum distance between every pair of centres of circles. We use a (100×100) grid.
2. e_2: How far the second biggest circle is from the mean of the other two, in terms of their radii. If they coincide then a score of 100 is given.
3. e_3: We do the same for this aspect of the visual structure, but instead of the radii of the circles we use their grey scale colour as our metric.

The characteristics show that a 'targeted' set of three circles with symetry for both radii and colour will be more 'ordered' or less 'disordered' rather than a set of circles with their centres at maximum distances between them, proportionally meaningless ratio of radii and values of colour not changing in a smooth fashion. The former should be mapped into a company that is performing really well on all aspect and the latter to a company performing badly on all aspects. Figure 1 shows an example of a 'healthy' looking company, whereas Figure 2 shows an example of an 'ill' performing one.

3.3.4 Step 4

The minimisation function is in this case:

$$\sum_{d=1}^{100} \sum_{s=1}^{3} (\nu_{ds} - e_{ds})^2$$

3.3.5 Step 5

This final step involved finding the suitable GP parameters for this problem. These parameters were chosen after empirical testing:

A population size of 750 running for 150 generations. The creation type is a ramped half-half as described by Koza [8], with a maximum depth of creation 7. Terminal set consists of the 8 variables shown in table 1, and function set consists of the four arithmetic operations $(+, -, \%, *)$ where % represents protected division. Protected division takes care of division by zero case. We used tournament selection for mating with a tournament size of 6. In generating subsequent populations 90% of the times we were using crossover, 5% of the times swap mutation and 5% of the times shrink mutation. We were always adding the best of the previous population to the next and we used a generational state GP.

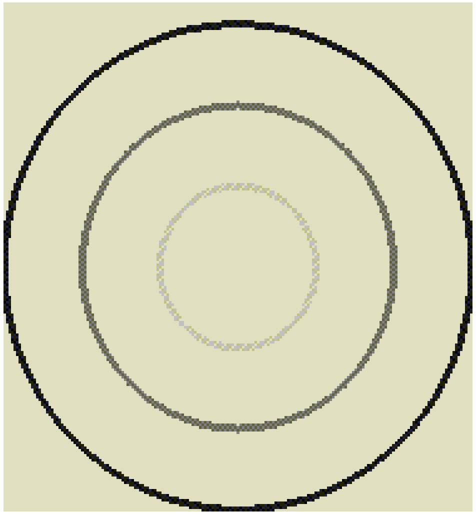

Figure 1: Picture of a "healthy" company.

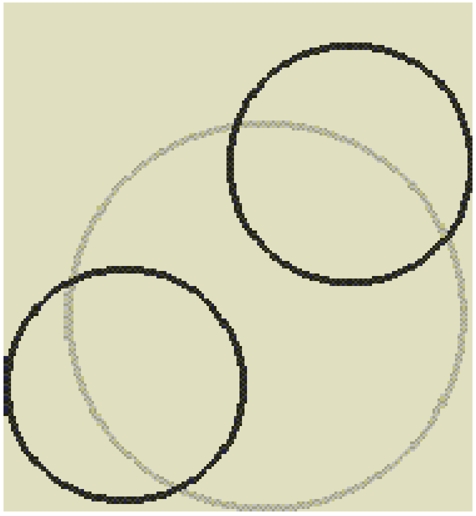

Figure 2: Picture of an "ill" company.

4 Results

Before we proceed with the analysis of the experiments we will show some results from the GP. Figure 3 shows, by generation, the standardised fitness of the best-of-generation individual and the worst-of-generation individual in the population for all generations $0 - 150$ of one run of this problem. The value for fitness is calculated out of equation 2. As can be seen, the fitness of the best-of-generation individual generally improves (i.e., decreases). The results shown in figure 3 are from the same run of the program that produced the visual structures for the experiments described below.

We ran 3 experiments with 10 subjects for each one of them. The subjects are postgraduate students of the University College London with no knowledge of the method prior to running the experiments.

In the first experiment, the subjects were asked to classify 60 visual structures into 3 clusters without any knowledge of how the visual structures were produced. The three clusters are:

1. Companies that are performing relatively well in terms of profitability, dividend payout and liquidity.

2. Companies that are performing relatively moderate in terms of profitability, dividend payout and liquidity.

3. Companies that are performing relatively badly in terms of all three above importance functions over the variables.

The visual structure as mentioned above is a set of three circles on a (100×100) grid, each circle having arbitrary radius and a grey scale colour.

The reason behind this experiment was whether the subjects would identify the qualitative measurements of the set of three circles without being told of what they were. Success in such experiment strengthness our case that we have chosen the right characteristics.

The second experiment involved performing the same classification all over again for the same data set. However, this time the subjects were actually told of the important characteristics of the visual structure were and how the mapping was produced.

In the final experiment we had 100 visual structures produced from a larger data set. Here, two more clusters were added making the total five.

4. Companies that are performing relatively well in two out of three 'value system' variables (profitability, liquidity and dividend payout) and moderate on the third.

5. Companies that are performing relatively well in one of the three important features of the data set and moderate on the other two.

Clearly, there is a only a slight difference and big overlap between some of the clusters (1-4 and 2-5 for example). One of the reasons of this experiment was to see if the subjects would recognise these subtle differences.

The simulated data was produced bearing in mind the five clusters above, with an equal number of companies in each cluster. Since there is an equal number of companies in each cluster there is an equal probability for a visual structure to belong to any of the three clusters. However, this information was not given to the subjects for obvious reasons.

We compare the performance of this method to what a random classification technique would produce. In the first two experiments we would expect a $\frac{1}{3}$ probability of guessing the right cluster from a random technique and therefore 20 'hits' (since there are 60 visual structures). A 'hit' represents the times the subject got it right. For the final experiment we would expect the probability of hits to be $\frac{1}{5}$, since we have 5 clusters with an equal probability of occuring. So, there should be 20 hits again since the number of visual structures for this experiment is a 100.

For the first experiment a hypothesis test was performed on the number of successes the subjects had against the expected number of successes that would have been randomly produced. We can confidently conclude (at 99.5% level) that the important aspects of the visual structures used can perform significantly better than a random technique, even when the users are not being told of those aspects. On average subjects got it right 34 times out of 60 so the successes sample mean is 0.57 against the 0.33 of the random technique.

The same hypothesis testing was performed for the second expriment. The results were improved further and from the new hypothesis testing we can confidently conclude (at 99.5% level) that the method performs significantly better than a random technique. This time subjects had on average 41 right responses out of 60 and therefore the sample mean has increased to 0.69. An example of the differences in the values is given in Figure 4.

For the final experiment we performed the same hypothesis test. In this experiment the success rate is lower than before to a mean of 0.46, or an average of 46 right guesses out of a 100. It is a clear consequence of the fact that it is hard to distinguish between clusters 1 and 4, or clusters 2 and 5. However, we can confidently conclude (at 99.5% level) that the method is still performing much better than a random technique expected to have a successes sample mean of 0.2.

5 Conclusions and Future Work

We have presented a new method for automatically mapping data to visual structures with the use of genetic programming techniques. This novel approach shows much promise. Questions like, "can subjects extract information from the visual representations of the data set?", or "can the visualisation act as an aid to the decision making process?", can be answered as yes with great confidence. Results also show that the method can be used as an initial step prior to statistical analysis. Maybe the first step to a classification device. Consider being an undergraduate admissions office or a bank offering loans to customers. You could use this method to produce a visual structure for each of your applicants and then trivial accepts and trivial rejects can be classified fast and with great of confidence. There will be applicants however that will need further investigation. As can be seen from the figure 4, this particular user always recognised the "ill" performing companies which in the case of financial data is very important since avoiding bad investments is critical.

As mentioned earlier, the project would like to investigate the potential of using natural representations for visual structures. Albeit, we used qualitative measurements over the set of circles, this particular visual structure is still very much abstract.

Therefore, the first step is to test the method with a more natural representation, the face. The familiarity of people with faces, the fact that is fast and easy to pick salient features and abnormalities and also the fact that a face evokes an emotional response to us, make us believe that the method will be improved further. Moreover the method will be tested with real data, and also tested for generalisability. What happens if, after the mapping is produced we add an extra row of similar data? Does the program still produce a 'meaningful' mapping?

Future work also involves, mathematically proving the method as valid. Furthermore, being able to query the visual structure and obtain more information about the data set, i.e. implementing a back-tracking technique for the genetic program. Moreover, testing the method with different data sets, different visual structures, different number of important characteristics and different methodologies to produce the mapping. Finally, there is the psychology aspect of the method. Being able to answer questions like, how "readable" the visual structure is.

6 Acknowledgements

Special thanks goes to the Knowledge Lab of NCR Financial Services who financially supports the early testing of this method.

References

[1] Marc Alexa and Wolfgang Muller. Visualisation by examples: Mapping data to visual representations using few correspondences. In *Proceedings of the Joint EUROGRAPHICS and IEEE TCVG, Symposium on Visualisation in Vienna, Austria, May 26–28, 1999*, pages 23–32. Springer-Verlag/Wien, 1999.

[2] J. Bertin. *Graphics and Graphic Information Processing.* deGruyter, 1981.

[3] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings In Information Visualisation: Using Vision To Think.* Morgan Kaufman, 1999.

[4] H. Chernoff. The use of faces to represent points in n-dimensional space graphically. Research Note NR-042-993, Department of Statistics, Stanford University, Gower Street, London WC1E 6BT, UK, december 1971.

[5] S. Feiner and C. Beshers. Worlds within worlds: metaphors for exploring n-dimensional virtual worlds. In ACM, editor, *UIST. Third Annual Symposium on User Interface Software and Technology. Proceedings of the ACM SIGGRAPH Symposium, Snowbird, Utah, USA, October 3–5, 1990*, pages 76–83, New York, NY 10036, USA, October 1990. ACM Press.

[6] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates for visualizing multi-dimensional geometry. In Tsiyasu L. Kunii, editor, *Computer Graphics 1987 (Proceedings of CG International '87)*, pages 25–44. Springer-Verlag, 1987.

[7] Walker J., Sproull L., and Subramani R. Using a human face in an interface. In *Human Factors in Computer Systems*, pages 85–91. ACM, 1994.

[8] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, 1992.

[9] Smith G. M. and Taffler R. Improving the communication of accounting information through cartoon graphics. In *Accounting, Auditing and Accountability Journal*, volume 9, pages 68–85, 1996.

[10] Ted Mihalisin, John Timlin, and John Schwegler. Visualizing multivariate functions, data, and distributions. *IEEE Computer Graphics and Applications*, 11(3):28–35, May 1991.

[11] Moriarity. Communicating financial information through multidimensional graphics. In *Journal of Accounting Research*, volume 17, pages 205–224, 1979.

[12] M.Slater. From data to visual structure: An automatic mapping. Research note, Department of Computer Science, University College London, 1999.

[13] Anselm Spoerri. Infocrystal: A visual tool for information retrieval. In *Proceedings IEEE Visualization'93*, pages 150–157. IEEE, 1993.

[14] Wilkinson. An experimental evaluation of multivariate graphical point representations. In *Human Factors in Computer Systems*, pages 202–209. ACM, 1982.

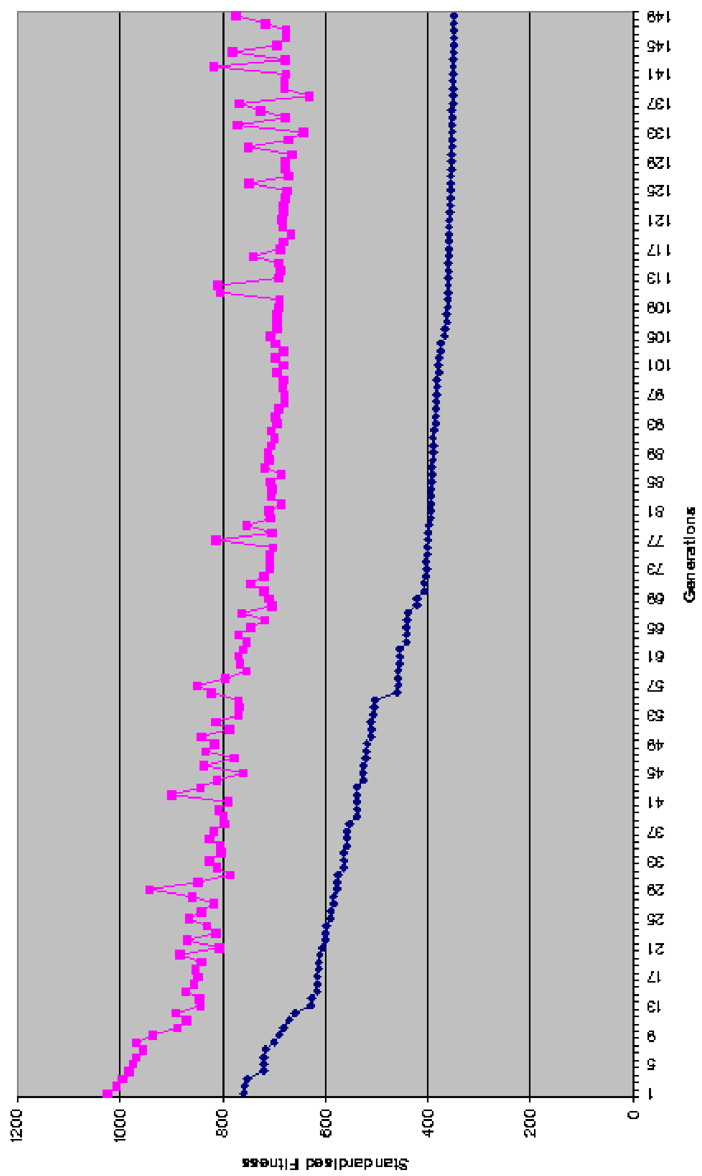

Figure 3: Fitness curves for one run of the GP.

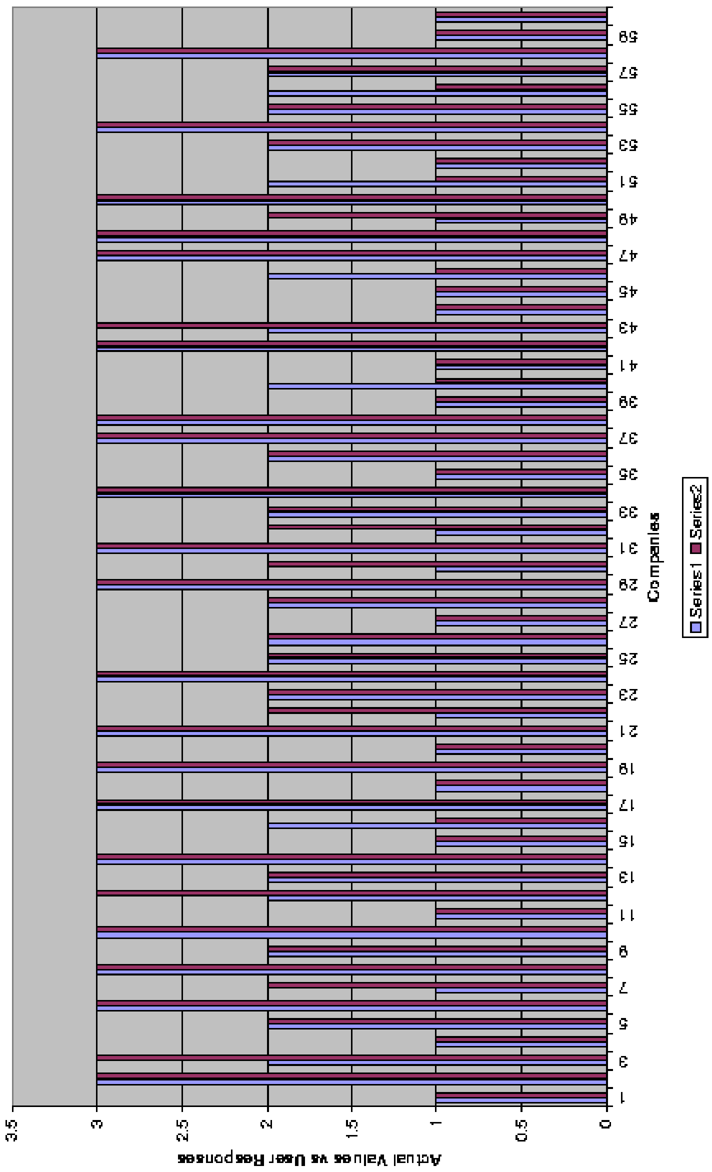

Figure 4: An example of actual data vs user responses from experiment 2.