EvoGraphDice : Interactive Evolution for Visual Analytics

Waldo Cancino, Nadia Boukhelifa and Evelyne Lutton
INRIA - Saclay-Ile-de-France, AVIZ team,
LRI - Bâtiment 650, Université Paris-Sud, 91405 ORSAY Cedex - France
http://www.aviz.fr
Email: waldo.cancino@inria.fr, nadia.boukhelifa@inria.fr, evelyne.lutton@inria.fr

***Abstract*—Visualization of large and complex datasets is a research challenge, especially in frameworks like industrial design, decision making and visual analytics. Interactive Evolution, used not only as an optimisation tool, but also as an exploration tool may provide some versatile solutions to this challenge. This paper presents an attempt in this direction with the EvoGraphDice prototype, developed on top of GraphDice, a general purpose visualization freeware for multidimensional visualization based on scatterplot matrices. EvoGraphDice interactively evolves compound additional dimensions, that provide new viewpoints on a multidimensional dataset. Compound dimensions are linear combination of the initial data dimensions, they are initialised with a Principal Component Analysis (PCA), and modified progressively by the interactive evolution process. Various interactions are available to the user, either in a transparent way, via a capture of mouse-clicks, or in a fully controlled manner, where the user has the opportunity to modify or include his own compound dimension in the evolved population, control the search space, or do some interactive queries. EvoGraphDice is tested on a synthetic dataset of dimension 6, where a known dependency is rediscovered via interactive manipulation. A second example is presented, based on a real dataset of dimension 13, provided by an industrial partner. Our experiments prove the potential of this interactive approach, and allow us to sketch future directions of development for the EvoGraphDice prototype.**

I. Introduction

Visual analytics can be defined as "the science of analytical reasoning facilitated by visual interactive interfaces" [1]. Research in this domain is trying to produce the most versatile and adapted tools for users to be able to visualize, understand, manipulate or transform multidimensional complex data for various tasks. In doing this, however, the designer of visual analytics systems is faced with the problem of finding the most appropriate methodological and parametric choices for the comfort and freedom of use. Additionally, the variety of end user needs, capabilities and cultural habits, make this problem even more complex. Consequently, current developed systems tend to provide more and more freedom and flexility to the user, potentially at the cost of overwhelming him with a maze of possible exploration paths.

Interactive Evolutionary Algorithms (IEA), i.e. optimisation algorithms that are able to optimise subjective data [2]–[5], may be an attractive solution to this paradox. The purpose of this paper, is to develop a prototype of interactive evolution aimed at helping users explore complex datasets. The idea is to let the IEA propose smaller sets of potential viewpoints for the user, based on past interactions.

There are many important issues to consider for proposing such a guiding system, among which: efficient and transparent use of interaction data, simultaneous visualization of different potential solutions, and management of diversity.

We feel that Interactive Evolutionary Algorithms (IEA) are convenient for guiding the user in a complex dataset without too many constrains. This opinion is founded by the following characteristics of simulated evolution:

- *focus:* by nature artificial evolution is a stochastic optimisation tool: IEA optimises a fitness function that aggregates some measures of comfort or satisfaction of the user, i.e. it has the ability to focus the attention of the user toward some "interesting" area of the search space,
- *diversity preservation:* diversity comes from two sources: *(a)* stochasticity of the elementary mechanisms, that usually makes EA robust for optimisation tasks, and *(b)* population-based scheme, that allows to display a variety of solutions to the user at any stage of the process,
  Interactive design, or exploration, is not only an optimisation task, diversity management needs special attention, and is an important component that deals with user's creativity [6], [7],
- *adaptation:* it is also known that EA are able to deal with varying environment; IEA have been proven to be able to follow change of user focus, see for instance [7].

Let us note, finally, that the focus of current research in interactive evolution seems to be more on industrial or artistic design tasks, than on data exploration tasks [8]. Our current work can be placed in the latter category and is strongly related to the field of diversity management [9].

The work described in this paper is part of the CSDL project (Complex Systems Design Lab, 2009-2012), a project funded by the French "System@tic pole" whose main contractor is an industrial partner, Dassault Aviation, and that involves 27 partners both from academy and industry. The objectives of the CSDL project is to set up a complete collaborative decision making framework for complex systems design purposes[1]. The role of our team in this project is to deal with the Visual Analytics issues. In the framework of computer simulation of very complex systems, interactive evolution is considered as a mean to :

[1] http://www.teratec.eu/activites/projetsR_D_Csdl.html

- improve exploration in very large parameter spaces, while controlling the computational power allocated to simulation calculations (refinement on demand for some areas of the parameter space),
- improve the understanding of some design compromises (Pareto front visualization),
- validate or reconsider intuitive or classical design choices,
- discover optimal and/or unconventional parameters setting.

Our prototype is based on a visualization tool developed in the AVIZ team of INRIA [10], GraphDice, to which we added evolutionary capabilities. The paper is organized as follows: related work in data visualization are presented in sections II-A and II-B, while previous work involving interactive evolution and visual analytics is discussed in section II-C. The prototype is presented in section III, and tested on two datasets (synthetic and real data) in section IV. Conclusions and future work are sketched in section V.

II. Related Work

A. Multidimensional Data Visualization and SPLOMs

Multidimensional visualization is concerned with providing visual tools to explore high-dimensional datasets and to reveal aspects of this data that are not, or at least not easily, revealed by standard statistical methods [11]. Here, each data attribute represents a dimension and thus the challenge is to map n-D data space onto a 2-D screen space. Beyond the problem of mapping and projections, there are issues of occlusion, dimension ordering and navigation. Many multidimensional visualization techniques have been proposed and classified; Keim [12] gives a taxonomy of primarily multidimensional visualization techniques. More specifically, Ward [13] describes a taxonomy for glyph placement strategies, and Valiati et al. [14] propose a task taxonomy for guiding evaluations of multidimensional visualizations.

In this paper, we use Scatter Plot Matrices (SPLOMs) [15], [16] as our primary multidimensional visualization method. SPLOMs, sometimes referred to as draftsman's plots, are well known for visualizing high-dimensional data and can be found in many statistical packages. Here, data dimensions are assigned to both rows and columns of a matrix; each cell in the matrix is a 2-D projection of the data depicted as a scatter plot. The method builds on user familiarity with scatter plots; known for their simplicity and high visual clarity [17]. The matrix itself is useful for quickly assessing bivariate combinations of interest. Thus, dependencies, clusters and outliers can be quickly identified. However, even with a small number of dimensions, scatter plot matrices can grow large in size resulting in thumbnail cells that are difficult to examine in detail. To deal with issues of high-dimensionality many systems offer SPLOM interactivity whereby, the overview matrix becomes the interface to drive the exploration of the 2-D bivariate space, and detailed views of the scatter plots are coordinated with the main display via brushing-and-linking [18], [19]. ScatterDice [20] is one such system but it differs from similar systems in that it offers a coherent and structured navigation through the n-dimensional space, as described in the next section.

B. ScatterDice and GraphDice Systems

ScatterDice [20] is an interactive visual tool for exploring multidimensional data sets. Users navigate the multidimensional space via simple 2-D projections, organised in a scatterplot matrix. This overview of thumbnail satterplots is linked to a detailed view where bivariate projections can be further investigated. One of the core features of the tool is that navigation in the multidimensional space, thus the transitions between scatterplots in the matrix, is performed as animated rotations in the 3-D space analogous to rolling a dice. Other features of the system include dimension reordering, visual queries and various interaction techniques for controlling transitions between scatterplots.

ScatterDice was further extended to support multivariate network visualization tasks, and the new system was called GraphDice [21]. It uses the main multidimensional navigation paradigm of ScatterDice but extends it with novel mechanisms to support network exploration in general and Social Network Analysis (SNA) tasks in particular.[2] We have chosen to work with GraphDice as our Visual Analytics (VA) tool for rapid prototyping and convenient access to expert users. GraphDice is an inhouse tool built by our research team and is already used by a number of research projects for visual exploration of multidimensional data sets (for instance, for visual analysis of evolutionary algorithms behavior, see [22]). The advantage is that users who are familiar with the GraphDice system, do not have to go through the steep learning curve of the new system.

C. Visual Analytics Meets Interactive Evolutionary Algorithms

Interactive Evolutionary Computation describes evolutionary computational models where humans, via suitable user interfaces, play an active role, implicitly or explicitly, in evaluating the outputs evolved by the evolutionary computation. Applications of IEC are varied ranging from art to science [4], [7], [23]. IEC lands itself very well to art applications such as for melody or graphic art generation where creativity is essential, due to the subjective nature of the fitness evaluation function. For scientific and engineering applications, IEC is interesting when the exact form of a more generalised fitness function is not known or is difficult to compute, say for producing a visual pattern that would interest a particular user. Here, the human visual system, together with the emotional and psychological responses of the user in question are far more superior than any pattern detection or learning algorithm.

Visual Analytics (VA) is a multidisciplinary field that integrates various sophisticated computational tools with innovative interactive techniques and visual representations to

[2] A demo of GraphDice can be launched from http://www.aviz.fr/graphdice/, it accepts standard .csv files (although it may be necessary to add a second line after the header giving the data type for each column - INT, STR, REAL, etc).

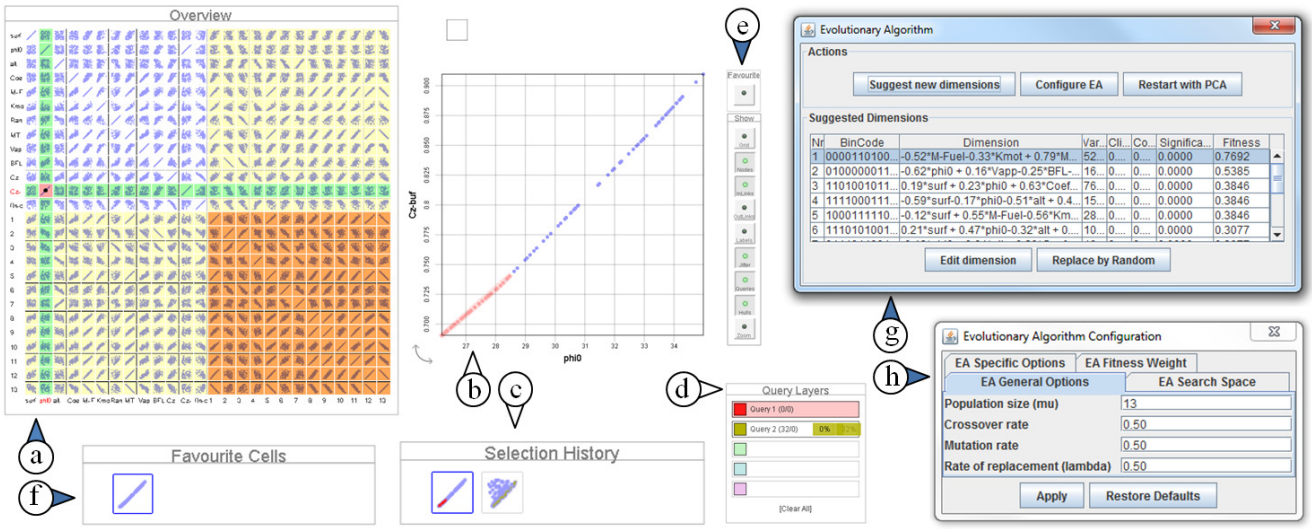


Fig. 1. EvoGraphDice prototype showing an exploration session of an aircraft simulation dataset: new extensions to the GraphDice system are indicated by coloured label arrows. (a) an overview scatter plot matrix showing the original data set of 13 dimensions and 13 new dimensions (numbered 1 to 13) as suggested by the evolutionary algorithm. (b) main plot view showing a linear correlation between two dimensions (weight and affinity) and a selected subset of points in red. (c) the selection history tool. (d) selection query window. (f) toolbar with "favourite" toggle button. (f) favourite cells manager stores interesting scatter plots. The user can bring a favourite cell back into the overview matrix. (g) the evolutionary algorithm main control window. (h) advanced configuration options for the evolutionary algorithm.

facilitate human interpretation of raw data [24]. Thus, VA can provide both the depiction techniques and the exploration tools necessary to help the user better evaluate the output of interactive evolutionary algorithms. Beyond works on visualizing the IEC search space [25], [26], to our knowledge, there has not been much work to marry efforts of the two communities; i.e. the VA and the IEC communities.

Moreover, despite efforts to design good user interfaces for IEC, human interaction with these systems, usually raises several problems, mainly linked to the "user bottleneck" [27], human fatigue and slowness. Various solutions have been considered [3], [27], [28] to deal with this issue, let us mention:

- the reduction of the size of the population and the number of generations, i.e. running a micro-EA,
- the use of specific models to constrain the research in *a priori* "interesting" areas of the search space,
- the use of a approximate user model (based on a limited number of characteristic quantities) that is used to filter obvious bad solutions and only present to the user the most interesting individuals of the population. The model can be learned, based on past interaction (see for instance [29]).

In what follows, we have chosen to run a micro-EA, with a population size constrained by the dimensionality of the dataset. Genetic diversity is thus an important issue, for avoiding uniformisation of the population (premature convergence) as well as user disinterest or fatigue: the algorithm has been tuned to display a set of diverse solutions at each generation. We will also see below that a variety of interactions are available with EvoGraphDice (query selection, direct interaction with the genome, search space modification). Multiple interactions with the evolutionary mechanisms have been proven to be very efficient in an artistic application [7], these ideas have been adapted here to the multidimensional data exploration task.

III. The EvoGraphDice prototype

We choose to run an Interactive EA on a population made of additional, compound dimensions, in order to match the structure of GraphDice that provides a variety of efficient and smooth visualization mechanisms.

A. Individual Initialization

The EA manipulates a population of dimensions of the same size to the number of variables in the data to be analysed (i.e. n individuals for n dimensions or variables).

Internally, the dimensions (i.e. the genomes) are represented as a linear combination of term, see (1), that allow the representation of linear combinations as well as more complex combinations of dimensions.

$$y = \sum_{i=1}^{n} w_i \times term_i + K, \tag{1}$$

where $term_i$ is a mathematical expression containing the i-th variable of the dataset (noted as x_i). $term_i$ represents raw, quadratic, logarithm, or exponential of x_i ($x_i, x_i^2, log(x_i), e^{x_i}$, respectively) or a combination of these expression (for instance $x_i \times log(x_i)$). The w_i refers to the weight of the $term_i$ and K is a constant. $term_i$ and K are real values.

The initial dimensions proposed by the EA should be interesting to the users in order to motivate them to explore the new alternatives. It was assumed that interesting solutions

should show relations among variables which are not evident in the original data.

In order to generate the initial dimensions, a PCA analysis [30] is performed on the original data. Thus, the n linear dependencies found by PCA conform the initial EA population.

These solutions are showed in EvoGraphDice as additional dimensions in the scatter plot matrix. Fig. 1(a) shows and example of the new dimensions suggested. Each dimension y_i found by the EA correspond to the $n+i$-th row and column of the scatter plot matrix. The additional n rows and columns are color highlighted in order to distinguish them from the ones belonging to the original data.

B. Fitness Evaluation

Evaluating the fitness of the suggested visualization requires taking into account user interactions and internal metrics. The user interaction criterion tries to adapt user preferences in the fitness function while the internal metrics evaluate the relations between variables. We use four criteria to build the fitness of each solutions: number of clicks, variance, complexity and significance.

All the interactions of the user with the SPLOM, both with the original dimensions (x_i) and the suggested ones (y_i), are recorded. Each time the user clicks on a cell of the scatter plot, the click counter of the corresponding original dimension (denoted by $freq(x_i)$) or additional dimension (denoted by $nclicks(y_i)$) is increased. These values are then normalized between 0 and 1.

The variance criterion, denoted by $vc(y_i)$, is based on the variance of the new dimension, as follows:

$$vc(y_i) = \frac{1}{1 + Var(y_i)}, \tag{2}$$

where $Var(y_i)$ is the variance of the dimension y_i. The term $vc(y_i)$ will prefer individuals with lesser variance and will mitigate the effects when working with variables of different scales. The criterion $vc(y_i)$ is intended to favour individuals (compound dimensions) that depict a dependency, i.e. corresponding to an approximation of an equation like $y_i = constant$

The complexity criterion, denoted by $comp(y_i)$, is related to the number of variables being represented in the genome:

$$comp(y_i) = \frac{\sum_{j=1}^{n} comp_j(y_i)}{n},$$
$$comp_j(y_i) = \begin{cases} 1, & w_j \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Finally, the significance criterion, denoted by $sig(y_i)$, is the sum of the frequencies of the variables that are represented in the terms of y_i, as can be seen in. (4):

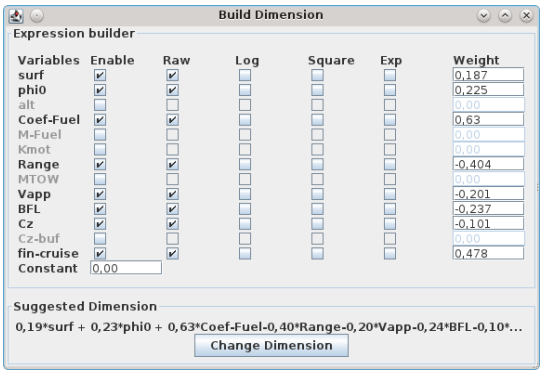


Fig. 2. EvoGraphDice suggestion dimension builder.

$$sig(y_i) = \sum_{j=1}^{n} sig_j(y_i),$$
$$sig_j(y_i) = \begin{cases} freq(x_j), & w_j \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The fitness of an individual results from the weighted sum of the four criteria described above as showed in (5):

$$fit(y_i) = W_1 \times nclicks(y_i) + W_2 \times vc(y_i) + W_3 \times (1 - comp(y_i)) + W_4 \times sig(y_i) \tag{5}$$

where the default weights of each criterion ($W_i, i = 1 \ldots 4$) are set to 1, although they can be adjusted by the user. The interactive part of the fitness function (significance $sig(y_i)$ and number of clicks $nclicks(y_i)$) rely on the hypothesis that the number of clicks on the cells are proportional to the interest of the user. This is of course a strong assumption, as the user may click for various reasons on a given cell other than for data exploration. Our experiments show that our current scheme is able to capture a reasonable approximation of the user's interest, enough to guide the evolution of the IEA.

New individuals are generated by tournament selection, barycentric crossover and Gaussian mutation. Additionally, a mutation that reduces the complexity of the expression is used: if the weight w_i of the i-the term is less than a weight threshold, w_i is set to zero. This way, an additional pressure toward low complexity individual is added.

C. Preserving Diversity

As the proposed EA deals with small population, the loss of diversity of solutions should be avoided. Each time a new dimension y'_i generated, the following conditions should satisfied:

- the distance between y'_i and all the individuals of the current population $y_i, i = 1 \ldots n$ should be greater than a threshold. We use the Euclidean distance between the weights of the y'_i terms and the weights from the y_i terms.
- the variance of the new individual is smaller than at least one of the individuals of the previous population.

While the first condition is important to maintain the diversity of the population, the latter condition allows to introduce new individuals with lower variance. If no solution satisfies these conditions, a random individual is generated. This operation is repeated, until λ new individuals replace the worst solutions of the previous population.

D. EvoGraphDice User Interface

EvoGraphDice also displays the current population as a table in a new window, see Fig. 1(g). Each line of this table corresponds to a compound dimension displayed in the SPLOM, columns give the identifier of the dimension (a number), a binary template that shows which primal dimension is used in the compound, the mathematical expression, the four fitness components (number of clicks, variance, complexity and significance) and the resulting fitness of each individual. The user can trigger an iteration of the algorithm ("Suggest new dimensions" button), configure the EA parameters ("Configure EA" button), restart the EA with a PCA analysis ("Restart EA with PCA" button), edit and alter any dimension ("Edit dimension" button) and replace the current dimension by a random one ("Replace by Random" button).

Dimensions can be edited using a simple expression builder window as showed in Fig. 2. The term expressions, their corresponding weights and the numerical constant can be changed at any time during EA execution. Moreover, EvoGraphDice allows to fully configure the relevant EA parameters as showed in Fig. 1(h). These parameters are:

- *crossover rate and mutation rate*, that sets the probabilities for the barycentric crossover, the Gaussian mutation and reduce complexity mutation operators,
- *replacement rate*, that refers to the proportion of individuals to be replaced at each EA iteration,
- *fitness threshold*, that corresponds to the minimal fitness difference between two individuals of the population. The individuals of a population are sorted according to their fitness. If the fitness difference between two successive individuals is smaller than the fitness threshold, both individuals are considered as equal. Then, the complexity of the two individuals are compared and the less complex one is ranked first,
- *weight threshold*, that defines the threshold under which the value of weights is set to 0, when applying the complexity reduction mutation,
- *sigma parameter*, the σ parameter of the Gaussian mutation,
- *fitness criteria weights*, that allow to tune the weights of each criterion (W_1, W_2, W_3, and W_4),
- *search variable and dimension space*, whose options are used to edit the expression associated to each variable. For instance, the user can associate the expression x_i^2 to the variable x_i such that all the individuals generated will have this term for the variable x_i. Moreover, the user can restrict the search to a subset of variable so the dimensions will contain only terms associated to the relevant variables,

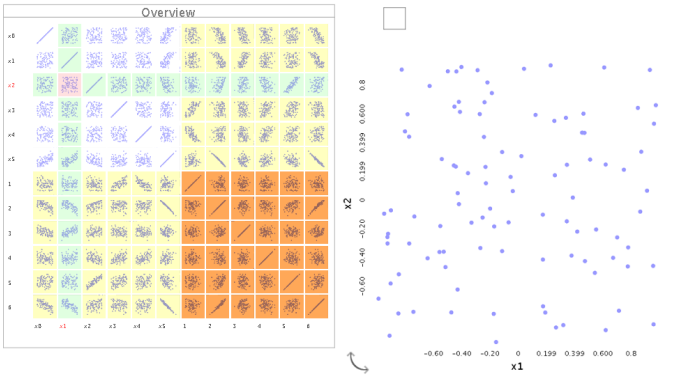


Fig. 4. Experiment with synthetic data : initial state.

- *work with query values*, that is used to restrict the set of points of the scatter plot on which the fitness is calculated. The search can thus be performed only on a subset of the data corresponding to a query set by the user.

Finally, the user can store the dimensions he feels are noteworthy or interesting in an "album" (the favorite cells widget, see Fig. 1(f). These dimensions can be re-inserted in the population at any time.

IV. Example of use on sample datasets

A. Visualization of a synthetic dataset

A first experiment has been run on a synthetic dataset, in which an interesting dependency has been set explicitly. The aim is to verify if the system is able to emerge dimensions that make the known dependency visually obvious.

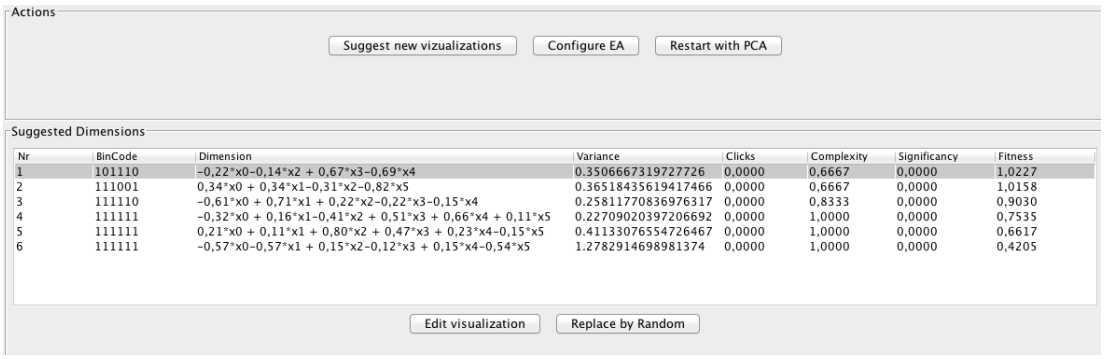
The data set is made of 100 points in 6 dimensions, the first 5 dimensions have been randomly generated in the interval $[-1, 1]$ with a uniform distribution, the 6th dimension is generated according to the following equation :

$$x_5 = \exp(x_0) + \exp(x_1) - x_2 \tag{6}$$

We thus have an artificial dataset that has two fully random dimensions (x_3 and x_4) and four dimensions that are linked by the above formula. This non-linear formula has been chosen on purpose, as linear dependencies would have been immediately found by the PCA initialization.

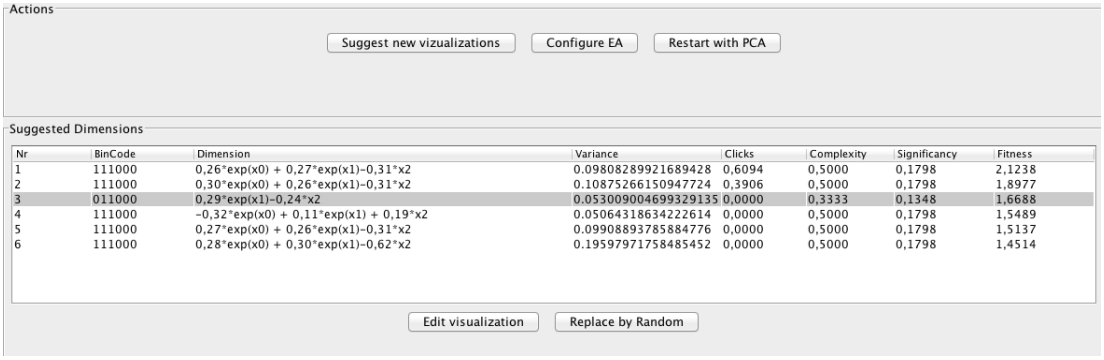
The initial state of Fig. 4 provides a PCA analysis, where no interesting feature is visible, even in the first compound dimension (tagged "1"), that corresponds to the lowest variance principal component. The population content is presented on Fig. 3.

Using our a priori knowledge on the dataset, we can decide to restrict the search space to compound dimensions that involves x_0 and x_1 as exponentials (see Fig. 8). Then in a dozen of generations, we get a dimension that resembles to the x_5 line of the scatter plot matrix (see Fig. 6). Fig. 5 shows that a dimension (the "1" dimension) has been evolved, making the dependency between x_0, x_1, x_2 and x_5 visually obvious. For comparison, Fig. 7 has been generated by artificially introducing the true dependency formula in the population.



Nr	BinCode	Dimension	Variance	Clicks	Complexity	Significancy	Fitness
1	101110	-0,22*x0-0,14*x2 + 0,67*x3-0,69*x4	0.3506667319727726	0,0000	0,6667	0,0000	1,0227
2	111001	0,34*x0 + 0,34*x1-0,31*x2-0,82*x5	0.36518435619417466	0,0000	0,6667	0,0000	1,0158
3	111110	-0,61*x0 + 0,71*x1 + 0,22*x2-0,22*x3-0,15*x4	0.25811770836976317	0,0000	0,8333	0,0000	0,9030
4	111111	-0,32*x0 + 0,16*x1-0,41*x2 + 0,51*x3 + 0,66*x4 + 0,11*x5	0.22709020397206692	0,0000	1,0000	0,0000	0,7535
5	111111	0,21*x0 + 0,11*x1 + 0,80*x2 + 0,47*x3 + 0,23*x4-0,15*x5	0.41133076554726467	0,0000	1,0000	0,0000	0,6617
6	111111	-0,57*x0-0,57*x1 + 0,15*x2-0,12*x3 + 0,15*x4-0,54*x5	1.2782914698981374	0,0000	1,0000	0,0000	0,4205

Fig. 3. Initial population content.



Nr	BinCode	Dimension	Variance	Clicks	Complexity	Significancy	Fitness
1	111000	0,26*exp(x0) + 0,27*exp(x1)-0,31*x2	0.09808289921689428	0,6094	0,5000	0,1798	2,1238
2	111000	0,30*exp(x0) + 0,26*exp(x1)-0,31*x2	0.10875266150947724	0,3906	0,5000	0,1798	1,8977
3	011000	0,29*exp(x1)-0,24*x2	0.053009004699329135	0,0000	0,3333	0,1348	1,6688
4	111000	-0,32*exp(x0) + 0,11*exp(x1) + 0,19*x2	0.05064318634222614	0,0000	0,5000	0,1798	1,5489
5	111000	0,27*exp(x0) + 0,26*exp(x1)-0,31*x2	0.09908893785884776	0,0000	0,5000	0,1798	1,5137
6	111000	0,28*exp(x0) + 0,30*exp(x1)-0,62*x2	0.19597971758485452	0,0000	0,5000	0,1798	1,4514

Fig. 5. Population content after a dozen of generations.

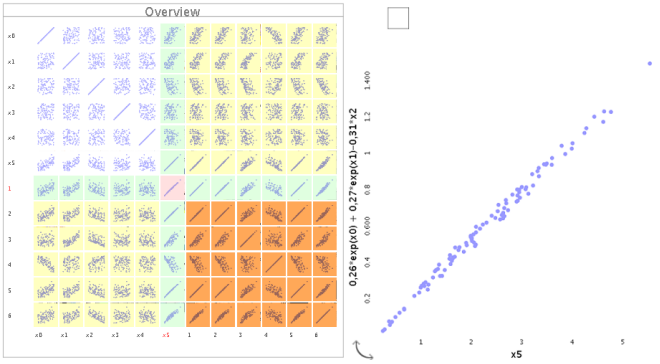


Fig. 6. Evolved state after a dozen of generations.

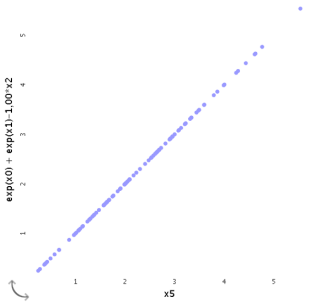


Fig. 7. Verification: visualization of the true formula by interactive modification of a genome.

B. Visualization of a real dataset

A second experiment has been ran on a real dataset[3], that represents simulation results of an aircraft model. In the dataset, displayed on Figs. 9–11, the 4 first parameters are input parameters ($surf$, $phi0$, alt, and *Coef-Fuel*) and the 9 left are output parameters. There are some obvious dependencies that are visible directly on the plain parameter view, like between $phi0$ and *Cz-buf* (see Fig. 1-b), and that exhibit two regimes (the dependency is linear by parts).

Our partner engineers are however interested in the behavior of parameters settings corresponding to Pareto frontiers, like for instance on Fig. 11 for compromises that minimise BFL while maximising the $Range$ parameter.

A first query has thus been sculpted on data corresponding to this Pareto set (in red on Fig. 11): it is used to restrict the fitness calculation. Additionally, the search has been restricted to formulas involving entry dimensions ($surf$, $phi0$, alt, and *Coef-Fuel*). The idea is to try to find if there may exist an approximative linear representation of the points of the Pareto front.

After a few experiments and a number of observations on the data, it has been noticed that a second query for low altitude points ($alt < 3900$, in green), that yields by intersection a subset of the previous Pareto front for low altitude, was enough

[3]Dataset courtesy of Dassault Systems (CSDL project partner).

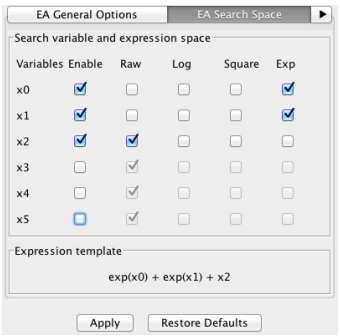


Fig. 8. Restriction of the search space to exponential values for x_0 and x_1

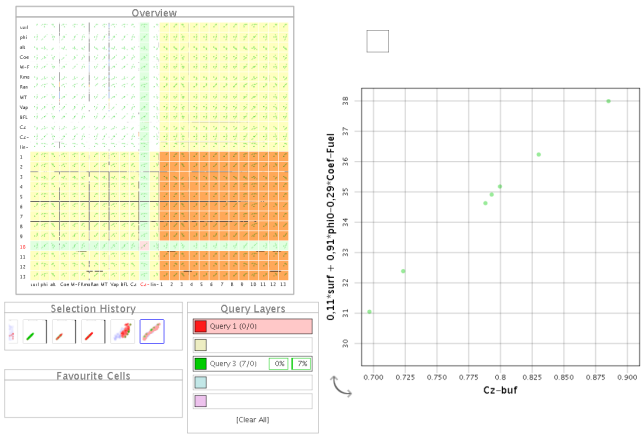


Fig. 9. Experiment with real data : detection of a linear dependence, visualization of the selected data only (green query).

to evolve in a few generations a dimension that provides a view where green points are aligned (see Fig. 9).

In this process, the red points have been used for the computed evaluation (with the aim of minising the variance), and the green ones are used for visual evaluation. A setting in GraphDice allows visualizing only the points of a query, via a transparency tuning for the unselected point (green points on Fig. 9). Fig. 10 gives the same view with all the points of the dataset, the Pareto front is in black.

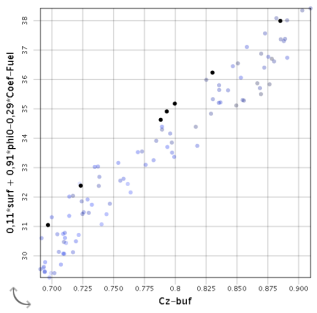


Fig. 10. Zoom on the evolved view : the black points are the one of the BFL-$Range$ Pareto set (Fig. 11).

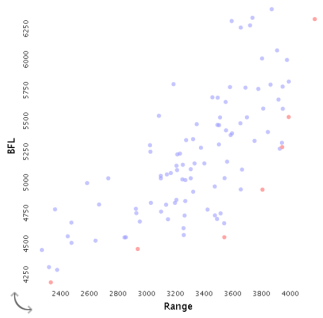


Fig. 11. Primary query, used for the calculation of the internal fitness: selection of the non-dominated points on the BLF-$Range$ view.

V. Conclusion and Future Work

The EvoGraphDice prototype has been experimentally shown to be able to evolve characteristics that are not visible in views based on the primary set of dimension (upper left quartile of the SPLOM matrix, see for instance Fig. 1), thus successfully fulfilling the main objective of this work. Additionally, no user-fatigue effects were reported during the experimental analysis, our intuition is that the variety of user interaction possibilities offered by GraphDice (i.e. not only related to the IEA process) is enough to maintain a good level of attention from the user.

A more precise evaluation of the added value due to interactive evolution will evidently need precise user-studies and statistical analysis. We intend to do this for future developments of the prototype. User-studies are important, no only for evaluation purposes but also to inform the default tuning of the internal parameters and strategies. The EvoGraphDice prototype is still in a preliminary stage; we currently propose ad-hoc tuning and give free access to parameter settings. Ideally, when addressed to people who do not currently practice EA, the system will need robust parameter settings, hidden away from the user. Decisions on default settings will be made on the basis of specific user studies focused on well defined characteristics, for instance, speed of emergence of an interesting view and diversity preservation.

Future developments of EvoGraphDice will also consider the following issues:

- preliminary experiments have shown the potential benefit of a more sophisticated encoding of compound dimensions: we intend to develop the next prototype on the basis of a Genetic Programming engine, in order to be able to smoothly evolve linear and non-linear compound dimensions within a single population,
- For the current version of EvoGraphDice we chose the number of mouse clicks as our metric for characterising user interest. We accounted for clicks on the SPLOM that result in a transition between views. Thus this metric corresponds more precisely to the number of times a cell was visited during an exploration session. Other metrics have been proposed in the literature; for example, in the context of web page browsing Claypool et al. [31] found

that visiting time was a good indicator of user's interest. We intend to carry out a user study to compare various implicit interest indicators with explicit user ratings in the context of visual analytics and multidimensional data exploration.

- the relationship between internal (computed) fitness and interactive fitness (capture of interaction events like mouse clicks) may be based on a learning step, in the same way as in [29], in order to be able to manipulate a large population, while displaying for interactive evaluation only the best individuals,
- engineer end-users were frustrated by the lack of an output interface, for instance with MatLab or SciLab, in order to be able to re-use the evolved dimension formulas, or to use the data subsets built by an interactive query. This function will be integrated in a future version of EvoGraphDice.

ACKNOWLEDGMENT

The authors would like to thank Michel Ravachol, from the Dassault Aviation company, for providing the dataset used in section IV-B , for his kind help, and fruitful discussions. We would also like to thank Jean-Daniel Fekete and Anastasia Bezerianos for their useful feedback and discussions.

REFERENCES

[1] J. Thomas and K. Cook, "A visual analytics agenda," *Computer Graphics and Applications, IEEE*, vol. 26, no. 1, pp. 10 – 13, jan.-feb. 2006.

[2] K. Sims, "Interactive evolution of dynamical systems," in *First European Conference on Artificial Life*, 1991, pp. 171–178, paris, December.

[3] W. Banzhaf, *Handbook of Evolutionary Computation*. Oxford University Press, 1997, ch. Interactive Evolution.

[4] H. Takagi, "Interactive evolutionary computation : System optimisation based on human subjective evaluation," in *IEEE Int. Conf. on Intelligent Engineering Systems (INES'98)*, Vienna, Austria, Sept 17-19 1998.

[5] E. Lutton, P. Grenier, and J. Lévy Véhel, "An interactive ea for multifractal bayesian denoising," in *EVOIASP*, 2005, lausanne.

[6] J. Chapuis and E. Lutton, "Artie-fract : Interactive evolution of fractals," in *4th International Conference on Generative Art*, Milano, Italy, December 12-14 2001.

[7] E. Lutton, "Evolution of fractal shapes for artists and designers," *IJAIT, International Journal of Artificial Intelligence Tools*, vol. 15, no. 4, pp. 651–672, 2006, special Issue on AI in Music and Art.

[8] M. Giacobini, A. Brabazon, S. Cagnoni, G. D. Caro, R. Drechsler, M. Farooq, A. Fink, E. Lutton, P. Machado, S. Minner, M. O'Neill, J. Romero, F. Rothlauf, G. Squillero, H. Takagi, S. Uyar, and S. Yang, Eds., *Applications of Evolutionary Computing*, ser. EvoWorkshops 2007: EvoCoMnet, EvoFIN, EvoIASP,EvoINTERACTION, EvoMUSART, EvoSTOC and EvoTransLog. Proceedings. Valence, Spain: Springer Verlag, April 11-13 2007, vol. LNCS 4448, http://www.springerlink.com/content/wltt81680351/.

[9] Y. Landrin-Schweitzer, P. Collet, and E. Lutton, "Introducing lateral thinking in search engines," *GPEM, Genetic Programming an Evolvable Hardware Journal, W. Banzhaf et al. Eds.*, vol. 1, no. 7, pp. 9–31, 2006.

[10] Niklas Elmqvist and Pierre Dragicevic and Jean-Daniel Fekete, "Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1141–1148, 2008, (Best Paper Award). [Online]. Available: {http://doi.ieeecomputersociety.org/10.1109/TVCG.2008.153}

[11] G. R. Loftus, "A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age." *Behavior Research Methods, Instruments, & Computers*, vol. 25, no. 2, pp. 250–256, 1993.

[12] D. A. Keim, "Information Visualization and Visual Data Mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[13] M. O. Ward, "A taxonomy of glyph placement strategies for multidimensional data visualization," *Information Visualization*, vol. 1, pp. 194–210, December 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=861129.861133

[14] E. R. A. Valiati, M. S. Pimenta, and C. M. D. S. Freitas, "A taxonomy of tasks for guiding the evaluation of multidimensional visualizations," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, ser. BELIV '06. New York, NY, USA: ACM, 2006, pp. 1–6. [Online]. Available: http://doi.acm.org/10.1145/1168149.1168169

[15] D. F. Andrews, "Plots of high-dimensional data," *Biometrics*, vol. 29, pp. 125–136, 1972.

[16] W. S. Cleveland and M. E. McGill, Eds., *Dynamic Graphics for Statistics*, ser. Statistics/Probability Series. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole, 1988.

[17] E. R. Tufte, *The Visual Display of Quantitative Information*. Graphics Pr; 2nd edition (May 2001), 1983.

[18] R. A. Becker and W. S. Cleveland, "Brushing scatterplots," *Technometrics*, vol. 29, no. 2, pp. 127–142, 1987.

[19] J. Seo and B. Shneiderman, "Rank-by-feature framework for interactive exploration of multidimensional data." *Information Visualization*, vol. 4, no. 2, pp. 99–113, 2005.

[20] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation," *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 2008)*, vol. 14, no. 6, pp. 1141–1148, 2008.

[21] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.-D. Fekete, "GraphDice: A System for Exploring Multivariate Social Networks," *Computer Graphics Forum (Proc. EuroVis 2010)*, vol. 29, no. 3, pp. 863–872, 2010.

[22] E. Lutton and J.-D. Fekete, "Visual Analytics of EA Data," in *Genetic and Evolutionary Computation Conference, GECCO 2011*, 2011, july 12-16, 2011, Dublin, Ireland.

[23] M. Fukumoto, S. Ogawa, S. Nakashima, and J. ichi Imai, "Extended interactive evolutionary computation using heart rate variability as fitness value for composing music chord progression," in *NaBIC*, 2010, pp. 407–412.

[24] J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0769523234

[25] N. Hayashida and H. Takagi, "Visualized IEC: interactive evolutionary computation with multidimensional data visualization," in *IECON 2000. 26th Annual Conference of the IEEE*, vol. 4, 2000, pp. 2738–2743.

[26] X. Llorà, K. Sastry, F. Alías, D. E. Goldberg, and M. Welge, "Analyzing active interactive genetic algorithms using visual analytics," in *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, ser. GECCO '06. New York, NY, USA: ACM, 2006, pp. 1417–1418. [Online]. Available: http://doi.acm.org/10.1145/1143997.1144223

[27] R. Poli and S. Cagnoni, "Genetic programming with user-driven selection: Experiments on the evolution of algorithms for image enhancement," in *Genetic Programming 1997: Proceedings of the Second Annual Conference*. Morgan Kaufmann, 1997, pp. 269–277.

[28] H. Takagi, "Interactive Evolutionary Computation: System Optimization Based on Human Subjective Evaluation," *INES'98*, 1998.

[29] E. Lutton, M. Pilz, and J. Lévy Véhel, "The fitness map scheme. application to interactive multifractal image denoising," in *CEC2005*. Edinburgh, UK: IEEE Congress on Evolutionary Computation, September, 2-5 2005.

[30] I. Smith, "A tutorial on principal component analysis," 2002.

[31] M. Claypool, P. Le, M. Waseda, and D. Brown, "Implicit interest indicators," in *In Intelligent User Interfaces*. ACM Press, 2001, pp. 33–40.