Computer Methods and Programs in Biomedicine (2005) **78**, 69—74

ELSEVIER

Computer Methods and
Programs in Biomedicine

www.intl.elsevierhealth.com/journals/cmpb

Augmented convex hull plots: Rationale, implementation in R and biomedical applications

Gaj Vidmar*, **Maja Pohar**

University of Ljubljana, Faculty of Medicine, Institute of Biomedical Informatics,
Vrazov trg 2, SI-1000 Ljubljana, Slovenia

Received 24 May 2004; received in revised form 20 December 2004; accepted 20 December 2004

KEYWORDS
Multivariate visualization;
Convex hulls;
Bivariate density;
R

Summary The paper addresses the possibility to replace cluttered multi-group scatter-plots with augmented convex hull plots. By replacing scatter-plot points with convex hulls, space is gained for visualization of descriptive statistics with error bars or confidence ellipses within the convex hulls. An informative addition to the plot is calculation of the area of convex hull divided by corresponding group size as a bivariate dispersion measure. Marginal distributions can be depicted on the sides of the main plot in established ways. Bivariate density plots might be used instead of convex hulls in the presence of outliers. Like any scatter-plot type visualization, the technique is not limited to raw data — points can be derived from any dimension reduction technique, or simple functions can be used as axes instead of original dimensions. The limited possibilities for producing such plots in existing software are surveyed, and our general and flexible implementation in R — the publicly available chplot function — is presented. Examples based on our daily biostatistical consulting practice illustrate the technique with various options.

1. Introduction

Convex hull drawing is a well-known computational geometry problem and there is a multitude of algorithms available for solving it. Even though links between computational geometry and statistics have been studied for quite a while [1—3], various tessellation methods are relatively rarely used in statistical visualization practice. An exception is the use of Voronoi diagrams in relation to discriminant analysis [4,5], while convex hull has been applied primarily for classification purposes in the field of pattern recognition [6—8].

Our paper focuses on the use of convex hulls rather than on the algorithms for finding them. There is a comprehensive resource with online Java implementation of different convex hull algorithms [9]. For solving the convex hull problem in *n* dimensions, the detailed *qhull* software package, which is based on the quickhull algorithm [10], has been developed and made publicly available [11].

In the next section, we present our concept of augmented convex hull plots and discuss its

* Corresponding author. Tel.: +386 1 5437783.
E-mail address: gaj.vidmar@mf.uni-lj.si (G. Vidmar).

0169-2607/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2004.12.003

elements. It should be stressed at the beginning that the basic idea of replacing, or at least accompanying cluttered scatterplots with some type of ''clouds'' is by no means novel. In the 1980s, convex hulls were presented as an aid to visualization of correspondence analysis results [12] and as such, found some application in the field of ecological ordination. At the same time, bivariate radial smoothers were discussed as one of the many possible "faces" one can superimpose on scatterplots [13], or even use instead of original points, although the authors considered that rather a last resort (judging by the fact that it is used in just one of the 20 figures in the article). Nevertheless, subsequent visualization research focused either on the choice of type and color of symbols for maximizing discrimination in scatter-plots [14], or on scatter-plots with smoothers [15], rather than on the idea which we elaborate.

2. Method

We believe that convex hulls are a suitable replacement for scatter-plots if the groups are large and there is considerable overlap of points between them. Our implementation is in two dimensions, but one can envision three-dimensional implementations, possibly incorporated into dynamic visualization in modern statistical, data-mining, and/or visualization software. The proposed convex hull plots are primarily intended for depicting differences between groups on continuous variables, possibly accompanying logistic regression, discriminant analysis, or some other classification technique.

By replacing scatter-plot points with a convex hull, one gains space for clear graphical presentation of descriptive statistics for both axes. For that purpose, we implemented basic "parametric" and "nonparametric" choices in terms of error-bar plots and confidence ellipses. Our choices for error-bar plot are the mean plus/minus multiple of standard deviation or standard error of the mean (i.e., the normal tolerance interval or the confidence interval for the mean), and the median with the interquartile range, while other variants might be various M-estimators and robust measures of dispersion. Alternatively, the user can choose to plot a confidence ellipse within the convex hull, thus also depicting correlation between variables. More complex diagrams depicting the whole distribution, such as notched boxplots or violin plots, should probably not be used as an augmentation to convex hulls because of visual clutter.

But since marginal distributions are of interest as such, we suggest the established way of plotting them separately, with corresponding line color or line style on the top and right-hand side of a two-dimensional plot, whereby either frequency polygons or kernel density plots are suitable choices if there are more than two groups. Another informative addition to the convex hull plot is the area of the convex hull divided by the group size, which is a dispersion measure related to mean absolute deviation in the sense that it weighs all deviations from the center equally. Instead of convex hulls, bivariate (trivariate in three dimensions) density plots can be used (with corresponding area/volume per point as dispersion measure), which should be preferred in the presence of outliers, whereby a relatively sparse grid (i.e., large bandwidth) produces results resembling convex hulls.

3. Implementation

3.1. Convex hull plotting in existing software

Among major commercial statistical software packages, SYSTAT® [16] is the only one that provides automated convex hull drawing. Convex hull is one of the optional additions to scatter-plots, together with related computational geometry procedures —– Voronoi tessellation, Delaunay triangulation, minimum spanning tree and travelling salesman path. Instead of the convex hull, one can also use a nonparametric kernel density estimator with user-specified probability. In the Version 10, which was available to us for testing purposes, multi-group convex hull plots can only be produced in two dimensions. There is a huge choice of univariate density plots that can be displayed on chart's borders, but it is not possible to superimpose error bars or boxplots on the main plot area. As a substitute, "Gaussian bivariate ellipses" can be placed on the plot. An example of such a plot for the famous Iris dataset [17] is provided in Fig. 1. However, possibilities for subsequently enhancing the plot are very limited due to the modest capabilities of the Graph window.

Automated convex hull drawing has also been implementated in Microsoft® Excel: it is included in the *cluster* add-in by Cinquegranni, who has developed an astonishing collection of publicly available statistical and/or visualization tools in Excel [18]. The convex hull drawing procedure is based on advice from Peltier [19], the leading Excel-charting

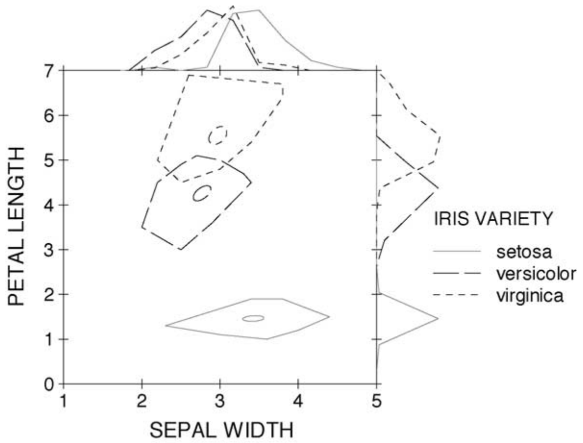

Fig. 1 A convex hull plot of the Iris dataset with confidence ellipses, produced with SYSTAT®.

expert. Although the procedure provides no data-selection or design-related options and just draws the convex hulls around the clusters determined by the clustering algorithm, the VBA code could be used for a more general implementation in line with of our ideas, which might be a worthy challenge for Excel-charting enthusiasts.

On the basis of an extensive search of bibliographic databases and the web, we conclude the list of existing convex hull implementation in statistical and related software with ADE-4 (acronym for Analyses des Données Écologiques), a system for "exploratory and euclidean methods in environmental sciences" [20]. Considering the French tradition of correspondence analysis [21] and ordination methods in ecology [22], it is no surprise that convex hull drawing has been implemented within it.

Because of platform independence, graphical power and flexibility, and because of compatibility of S code [23,24] with the widespread commercial S-PLUS® system [25], we opted for implementation of our ideas in the freely available R system [26]. Our starting point was the *chull* function for convex polygons in two dimensions, which was ported from S-PLUS® to R in 1999 [27]. Recently, convex hulls have been applied within the *multiv* package [28] in the *hierclust* function as the default option for representing clusters in "movie mode", and the ADE-4 system has been ported to R [29], but convenient production of publication-quality convex hull or bivariate density plots with freely chosen data and various options for additional information and marginal distributions call for a self-standing function. Hence, we developed the *chplot* function, which was later expanded into the *chplot* package.

```
chplot (x, y, groups, chull=TRUE,
  clevel=0.95, band.power=0.2,
  mar.den=FALSE,
  descriptives=‘‘mean.sd’’,
  dlevel=0.68, bw=FALSE,
  ratio=0.75, plot.points=FALSE,
  log=‘‘’’, xlab, ylab,
  legend.control (include=TRUE,
  area.in, pos, cex, bty, title),
  ...)

chadd (param, pos, add.fun, ...)
```

3.2. Our implementation in R: the *chplot* package

The package consists of four components: the *chplot* function, the *legend.control* object for detailed legend control, the *chadd* function that enables the user to freely add further elements to the augmented convex hull plot, and the *hdr* dataset presented in the next section. Detailed documentation is included in the package, so here we just list the syntax, which is self-explanatory to a large extent, and give an overview of the functions.

The *chplot* function requires just three vectors of data as input (*x*, *y* and group code), but offers numerous optional parameters (with the most generally useful values as defaults) and retains all the flexibility of R's plotting routines, so that the user can produce a wide variety of plots with precisely controlled features. The first choice for the user to make is whether to draw convex hulls, which is the default, or bivariate density contours. In the latter case, confidence level can be set (the default is 95%) as well as bandwidth (based on [30], the default value is group-size$^{-0.2}$). Next, marginal density plots can be chosen for plotting marginal distributions, whereby the same density scale is automatically used for both dimensions. The default option produces relative frequency polygons with points in the middle of the intervals and the minimum and maximum interval with zero frequency depicted for all the groups on a common relative frequency scale.

The default option for depicting descriptive statistics within convex hulls, or density contours produces a cross with the lines intersecting at the mean of *x* and *y* for each group and depicting the 68% tolerance interval (i.e., stretching one standard deviation in both horizontal and vertical directions). Standard error of the mean can be chosen instead of the standard deviation; specifying zero-level plots empty convex hulls or density contours, regardless of the variability measure. A non-symmetric alternative is to make the lines depict the first and the third quartile for both axes and cross at the median of *x* and *y*. Instead of the crosses, which are always parallel to the axes of the main plot, confidence ellipses can be selected, which are inclined in accordance with the correlation between *x* and *y* for the given group centered at the respective means and depict the bivariate confidence interval for the mean.

The whole plot can be produced either in color, which is the default, or in black-and-white, in which case a different line pattern is used for each group. Next, the user can specify the ratio of the main plot to the total plot area, which also determines the default legend placement: if the ratio is less or equal to the default value of 0.75, the legend is placed in the top right corner, otherwise it is placed within the main plot and positioned by the user with mouse. If the ratio is 1, marginal distributions are not plotted (and the legend is, of course, placed inside the plot). On the other hand, the user can make the plot more crowded by choosing to plot all the raw data points in addition to the convex hulls or density contours.

The legend can be omitted; if present, as it is by default, its entries are group names (i.e., factor level names in R). If area per point has been calculated, it is reported after group name in parentheses. The area is calculated by default for convex hulls, while for bivariate density plots it can be requested. When the density contours are non-contiguous or even just line fragments, the area per point does not make any sense, and the confidence level and/or bandwidth formula power should be adjusted. One or both variables can be plotted on logarithmic scale; the setting applies to the main plot as well as to the marginal plot(s). If desired, axis titles different from *x* and *y* variable names can be set.

Default legend font size can be overridden, and the same goes for legend frame (which is omitted by default if the legend is placed outside the main plot). When the legend is outside the main plot, the legend title can be set. The ellipsis symbol at the end of the *chplot* function syntax indicates that further arguments related to the legend can be specified in accordance with the R's *plot* function.

Three more characteristics of the *chplot* function are noteworthy: first, cases with missing value on any of the three compulsory variables are excluded from the plot (the function displays a warning to that effect); second, instead of specifying three vectors, only one data-frame (or matrix) with two or three columns can be specified as *x* (the second column being treated as *y* and the third one containing group membership), which makes variable names (if they exist) automatically appear as axis labels and in the legend; third, after producing the plot, the function restores all the graphics parameters to their previous values.

In addition to drawing the plot, the *chplot* function also returns a list object, which is used by the *chadd* function. This function adds crucial functionality, since it allows the user to freely add further elements (from boxes and lines to complex plot) to the augmented convex hull plot.

3.3. Availability

The package has been included in CRAN [31], from where the package source, binaries and reference manual can be downloaded [32]. It was introduced in R version 1.9. The package depends on the *KernSmooth* package [33], since it uses the *bkde2D* function for bivariate kernel density estimation, and on the *ellipse* package [34] for drawing ellipses.

4. Examples

We demonstrate the technique and the capabilities of our software with two datasets from our daily consulting practice. The first one comes from a large collaborative research project on socio-economic determinants of mortality in Slovenia, the results of which have partly been published in the UNDP and government sponsored report on human development and health [35]. The data were obtained by combining the population register (which contains all the census data), the electronic records of compulsory death registration forms and the personal income tax database. The *hdr* dataset contains data on the deceased in 1998 for whom the amount of personal income tax paid could be identified ($N = 9051$). Since missing data issues have not been resolved with the providers yet, the dataset cannot be considered representative of the entire population, but it is very useful for demonstrative purposes. Fig. 2 depicts gender differences in age at death and income tax with convex hulls, whereby the general trend of women living somewhat longer and earning somewhat less than men is clear. Logarithmic axis is chosen for income because the distribution is heavily right-skewed; descriptive statistics (mean ± standard deviation) are depicted with error bars, marginal distributions are plotted with relative frequency polygons, and information on area of convex hull per point is

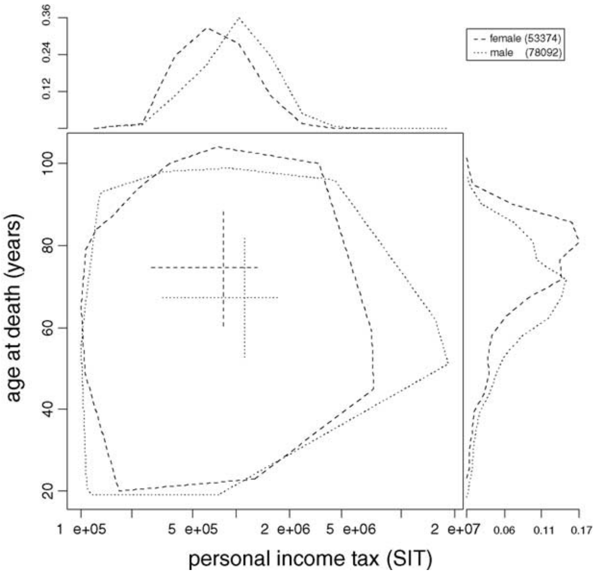

Fig. 2 Augmented convex hull plot of gender differences in age at death and paid personal income tax for deceased in Slovenia in 1998 [35]. R syntax: chplot(hdr, log=‘‘*x*’’, bw=TRUE, *x*lab=‘‘personal income tax (SIT)’’, *y*lab=‘àge at death (years)’’, title=FALSE, bty=‘‘0’’).

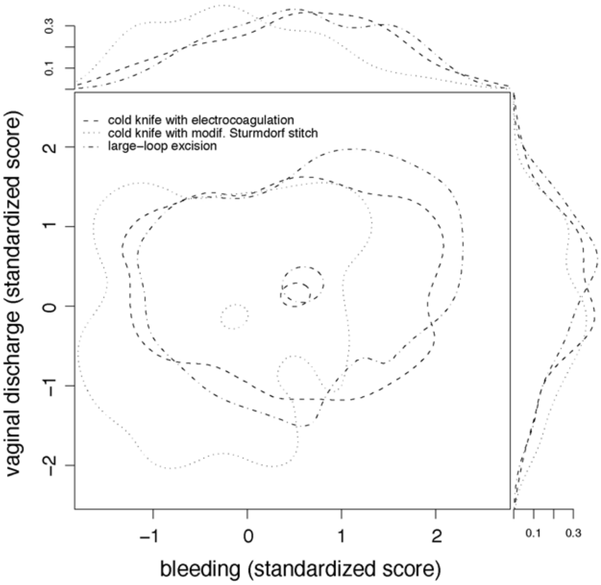

Fig. 3 Augmented bivariate density plot of standardized score of the number of days with patient-reported bleeding and/or vaginal discharge in relation to conization technique [36]. R syntax: chplot(conization, chull=FALSE, mar.den=TRUE, descriptives=‘èllipse’’, bw=TRUE, pos=‘ìn’’, bty=‘‘*n*’’, cex=0.7).

added (indicating greater variability among men); due to the default main plot-to-total ratio of 0.75, the legend is placed outside the main plot.

The second example is from the field of gynecology, from a study of risk factors for complications after conization [36] in which post-operative outcome was assessed with several "objective" (e.g. increased body temperature) and "subjective" measures (e.g. self-assessed pain) in over 800 women. In Fig. 3, two derived outcome measures (normalized and rescaled number of post-operative days with the patient reporting bleeding and/or vaginal discharge) are plotted for different operation techniques. Descriptive statistics are depicted with confidence ellipses, illustrating the fact that the two chosen dimensions are relatively independent ($r = -0.15$); marginal distributions are plotted with density plots, and since the ratio of the main plot to the total plot area is 0.9, the legend was placed inside the main plot by default and then positioned interactively. The plot indicates what has been confirmed by various multivariate models, namely that one technique (cold-knife conization with modified Sturmdorf stitch) is superior to the other two in terms of fewer post-operative complications.

Of course, the convex hull technique — like any scatter-plot type visualization — is not limited to "raw" data values. As already stressed in the introduction, sets of points can be derived from any multivariate dimension reduction technique, e.g. factor analysis, principal component analysis, multidimensional scaling [37], or correspondence analysis [38], or any functions can be used as axes instead of original dimensions.

5. Conclusion

The proposed augmented convex hull plot with all its variants represents just one of the many possibilities of using computational geometry algorithms for statistical visualization. In that sense, it is at least remotely related to such important and elegant concepts and methods as Voronoi and Delaunay tessellations.

Our basic idea was to eliminate visual clutter from multi-group scatter-plots of large datasets by replacing points with convex hulls or bivariate density contours. That brings about many possibilities for enhancing the exposition of the data: descriptive statistics can be plotted with error bars or confidence ellipses within the convex hulls, a bivariate dispersion measure (area of convex hull or density polygon per point) can be added to the legend, and marginal distributions can be plotted to the side and above the main plot with frequency polygons or density curves.

All these options are easily invoked and controlled with an R function named *chplot*, relieving the user from tedious manual tuning and assuring a coherent look of the chart. Together with the *legend.control* object for detailed legend control, the *chadd* function that enables the user to freely add further elements to the augmented convex hull plot and the *hdr* dataset; it has been included in the *chplot* package in order to make it properly documented and easilly installable through R's packaging mechanism. The package is publicly available at CRAN [31].

Augmented convex hull plots were designed for reducing visual clutter, but as the number of groups increases, even they eventually become unclear. Nevertheless, it is our experience that up to 10 groups can be nicely plotted with the majority of real-life datasets. Even though the expert opinion is that implementation of a 3D convex hull algorithm in R is a difficult task [39], a possibility for the future might be the extension of the procedure to three dimensions, possibly using the *Scatterplot3d* package [40], or even making the visualizations interactive with the *rgl* package [41]. However, as with any newly proposed visualization type, the immediate challenge is to see whether augmented convex hull plots gain wider acceptance.

References

[1] B. Efron, The convex hull of a random set of points, Biometrika 52 (1965) 331—453.

[2] M. Shamos, Geometry and statistics: Problems at the interface, in: J. Traub (Ed.), Recent Results and New Directions in Algorithms and Complexity, Academic Press, New York, 1976, pp. 251—280.

[3] B. Ripley, Spatial Statistics, first ed., Wiley, New York, 1981.

[4] H. Kamiya, A. Takemura, On rankings generated by pairwise linear discriminant analysis of *m* populations, J. Multivariate Anal. 61 (1997) 1—28.

[5] G. Ragozini, A data-driven discriminant rule by Voronoi tessellation, NTTS '98—Seminar on New Techniques and Technologies for Statistics: 4—6 November 1998, Sorrento, Italy, 1998, http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98/papers/cp/071c.pdf.

[6] Y.U. Degytar, M.Y. Finkelshtein, Classification algorithms based on construction of convex hulls of sets, Eng. Cybermetics 12 (1974) 150—154.

[7] V. Di Gesu, M.C. Maccarone, Description of fuzzy images by convex hull technique, Proceedings of the 8th International Conference on Pattern Recognition (ICPR), International Association for Pattern Recognition, Paris, 1986, pp. 1276—1278.

[8] L.Y. Shan, M. Thonnat, Description of object shapes by apparent boundary and convex-hull, Pattern Recognition 26 (1993) 95—107.

[9] http://www.cse.unsw.edu.au/∼lambert/java/3d/hull.html.

[10] C.B. Barber, D.P. Dobkin, H. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Trans. Mathematical Software (TOMS) 22 (1996) 469—483.

[11] http://www.thesa.com/software/qhull/.

[12] M.J. Greenacre, Theory and Applications of Correspondence Analysis, first ed., Academic Press, London, 1984.

[13] W.S. Cleveland, R. McGill, The many faces of a scatterplot, J. Am. Stat. Assoc. 79 (1984) 807—812.

[14] S. Lewandowsky, I. Spence, Discriminating strata in scatterplots, J. Am. Stat. Assoc. 84 (1989) 682—688.

[15] W.S. Cleveland, Visualizing Data, first ed., Hobart Press, Summit, 1993.

[16] http://www.systat.com.

[17] R.A. Fisher, The use of multiple measurements in taxonomic problem, Ann. Eugenics Part II (1936) 179—188.

[18] http://www.prodomosua.it.

[19] http://peltiertech.com.

[20] J. Thioulouse, D. Chessel, S. Dolédec, J.M. Olivier, ADE-4: a multivariate analysis and graphical display software, Statistics Computing 7 (1997) 75—83.

[21] J.-P. Benzecri, L'analyse des données tome 2: l'analyse des correspondances, Bordas, Paris, 1980.

[22] P. Legendre, L. Legendre, Numerical Ecology, second ed., Elsevier, Amsterdam, 1998 (in English).

[23] R.A. Becker, J.M. Chambers, A.R. Wilks, The New S Language, Chapman & Hall, London, 1988.

[24] J.M. Chambers, Programming with Data: A Guide to the S Language, Springer, New York, NY, 1998.

[25] http://www.insightful.com/products/splus.

[26] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2004, ISBN 3-900051-00-3, http://www.R-project.org.

[27] http://maths.newcastle.edu.au/∼rking/R/devel/99a/0299.htm.

[28] http://cran.r-project.org/doc/packages/multiv.pdf.

[29] http://cran.r-project.org/doc/packages/ade4.pdf.

[30] B. Everitt, S. Rabe-Hesketh, Analyzing Medical Data using S-PLUS, Springer, New York, NY, 2001.

[31] http://cran.r-project.org/.

[32] http://cran.r-project.org/src/contrib/Descriptions/chplot.html.

[33] http://cran.r-project.org/doc/packages/KernSmooth.pdf.

[34] http://cran.r-project.org/doc/packages/ellipse.pdf.

[35] J. Javornik, V. Korosec (Eds.), Human development report Slovenia 2002/2003: Human development and health, Institute of Macroecomic Analysis and Development, Ljubljana, 2003.

[36] A. Zupancic-Pridgar, Influence of vaginal flora on morbidity after conization, M.Sc. thesis, University of Ljubljana, Faculty of Medicine, Ljubljana, 2003.

[37] E.D. Gallagher, D. Shull, Statistical analyses of Boston Harbor benthos: 1991—1998, University of Massachusetts, Department of Environmental, Coastal and Ocean Sciences, Boston, 1999, http://www.es.umb.edu/faculty/edg/files/pub/bh98rept2.pdf.

[38] T. Pipan, Ecology of copepods (Crustacea: Copepoda) in percolation water of the selected karst caves, Ph.D thesis, University of Ljubljana, Faculty of Biotechnology, Department of Biology, Ljubljana, 2003.

[39] http://maths.newcastle.edu.au/∼rking/R/help/01c/1404.html.

[40] U. Ligges, M. Mächler, Scatterplot3d—an R package for visualizing multivariate data, J. Stat. Software 8 (11) (2003), http://www.jstatsoft.org/v08/i11/JSSs3d.pdf.

[41] D. Adler, RGL: 3D visualization device system (OpenGL), http://wsopuppenkiste.wiso.uni-goettingen.de/∼dadler/rgl/.