Communications of the Korean Statistical Society
2009, Vol. 16, No. 1, 185–193

Visualizing Multi-Variable Prediction Functions by Segmented k-CPG's

Myung-Hoe Huh[1,a]

[a]Dept. of Statistics, Korea Univ.

Abstract

Machine learning methods such as support vector machines and random forests yield nonparametric prediction functions of the form $y = f(x_1, \ldots, x_p)$. As a sequel to the previous article (Huh and Lee, 2008) for visualizing nonparametric functions, I propose more sensible graphs for visualizing $y = f(x_1, \ldots, x_p)$ herein which has two clear advantages over the previous simple graphs. New graphs will show a small number of prototype curves of $f(x_1, \ldots, x_{j-1}, x_j, x_{j+1} \ldots, x_p)$, revealing statistically plausible portion over the interval of x_j which changes with $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$. To complement the visual display, matching importance measures for each of p predictor variables are produced. The proposed graphs and importance measures are validated in simulated settings and demonstrated for an environmental study.

Keywords: Visualization of prediction functions, k-Means clustering, variable importance, support vector machine, random forests, environmental data.

1. Introduction

Obtaining prediction functions of the form $y = f(x_1, \ldots, x_p)$ by building a supervised learning model from a dataset of n observations of the response y and p predictors $x_1, \ldots, x_p$, we may ask ourselves what the function f looks like? If the assumed model allows non-additivity in predictors, the visualization task is absolutely non-trivial for $p \geq 3$. In previous article (Huh and Lee, 2008), the authors proposed to draw n conditional predictive graphs (CPG's) for each of p predictors, where each CPG(j), $j = 1, \ldots, p$ consists of trajectory curves

$$(t, f(x_{i1}, \ldots, t, \ldots, x_{ip})), \quad \text{for all } t \text{ in } (a_j, b_j),\ i = 1, \ldots, n.$$

Supporting intervals $(a_1, b_1), \ldots, (a_p, b_p)$ for $x_1, \ldots, x_p$ can be determined either by principal researchers or simply from the data, for instance

$$a_j = \min_{i=1,\ldots,n} x_{ij}, \quad b_j = \max_{i=1,\ldots,n} x_{ij}, \quad \text{for } j = 1, \ldots, p.$$

As illustration of the method, I draw CPG's of a support vector machine classifier for the `iris` subset data($p = 4$) in which the species are restricted to `versicolor` and `virginica`($n = 100$). See Figure 1 (reproduced from Figure 4.2 of Huh and Lee (2008)).

Each panel of CPG's in Figure 1 contains $n(= 100)$ curves over the same interval, ranging from the minimum to the maximum of respective variables. But, it would be better if the supporting interval (a_j, b_j) of x_j varies depending on the conditioning point $\mathbf{x}_{(j)} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$. Also,

[1] Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr.

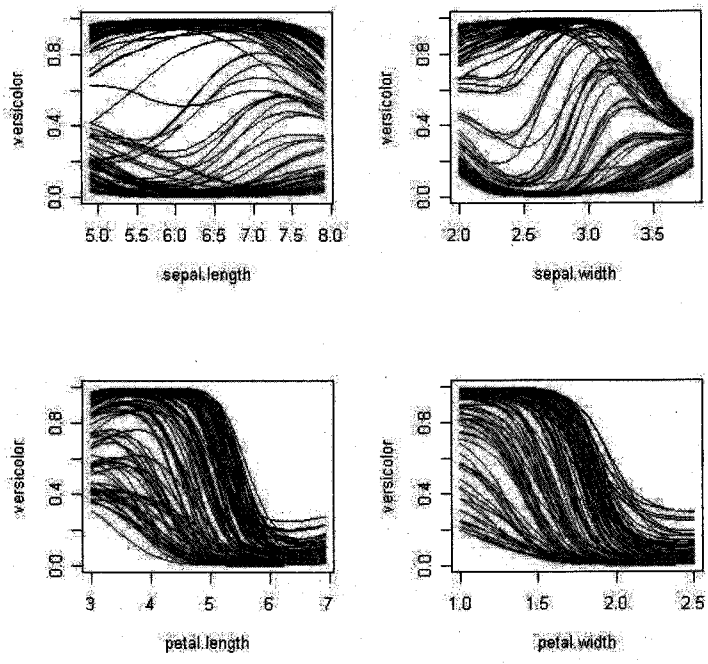


Figure 1: *CPG's of support vector machine classifier for* `iris` *data,* `Versicolor` *vs.* `Virginica`

previous study (Huh and Lee, 2008) revealed that the plot of curves looks cumbersome when n is somewhat large. The aim of this study is to develop an enhanced version of CPG's for visualizing $y = f(x_1, \ldots, x_p)$, produced by machine learners.

In Section 2 of this study, I propose more sensible graphs for visualizing $y = f(x_1, \ldots, x_p)$, herein new graphs will show a batch of k trajectories

$$\left\{(t,\ f(x_{i1}, \ldots, t, \ldots, x_{ip})) \,|\, t \in C(x_1, \ldots, x_{j-1},\ x_{j+1}, \ldots, x_p)\right\}$$

over chosen segments $C(x_1, \ldots, x_{j-1},\ x_{j+1} \ldots, x_p)$ for x_j. To advantage of the new graphs, importance measures for predictors matching to the graphs will be available. In Section 3, the proposed graphs and importance measures are validated under the simulated settings. In Section 4, the proposed graphs and importance measures are demonstrated for an environmental study of the ozone, modeled by the support vector machine and the rain forests. Finally, in Section 5, I conclude the article with several remarks.

2. k Prototype Curves on Plausible Intervals

The CPG(j) contains n curves($j = 1, \ldots, p$), one for each conditioning variate $\mathbf{x}_{i(j)} = (x_{i1}, \ldots, x_{i,j-1}, x_{i,j+1} \ldots, x_{ip})$, $i = 1, \ldots, n$. Even for moderately large n such as 400, the CPG's are almost darkened. To avoid such problematics, one may plot a fraction of curves (as illustrated in Figure 3.2 of the previous study). Another alternative is to use a manageable number of prototype conditioning variates for each $j = 1, \ldots, p$. I will pursue the latter approach employing k-means clustering to reduce the number of curves. K-means clustering is regarded as one of the standard methods for selection of prototypes (Hastie *et al.*, 2001).

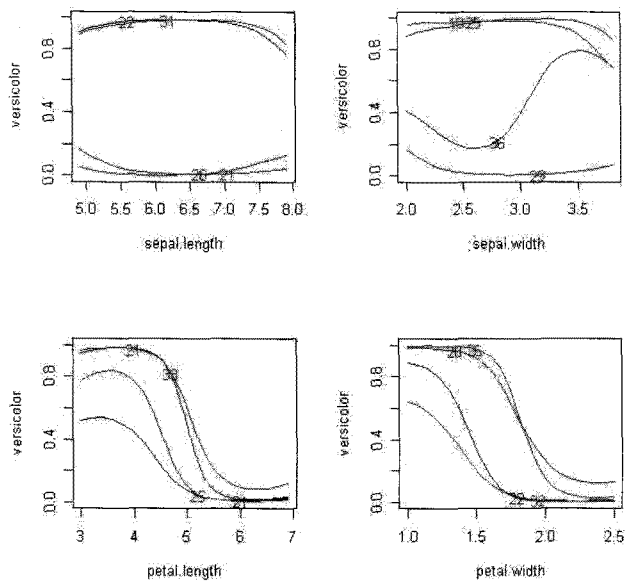


Figure 2: k-CPG's of SVM classifier for `iris` subset data, `Versicolor` vs. `Virginica`

k-CPG's: The First Proposal

1) For each $j = 1, \ldots, p$, reduce n conditioning variates

$$\mathbf{x}_{i(j)} = \left(x_{i1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{ip}\right), \quad i = 1, \ldots, n$$

to k prototype conditioning variates by k-means clustering which yields

$$\mathbf{x}^0_{l(j)} = \left(x^0_{l1}, \ldots, x^0_{l,j-1}, x^0_{l,j+1}, \ldots, x^0_{lp}\right), \quad l = 1, \ldots, k.$$

2) Plot trajectory curves

$$\left(t,\, f\left(x^0_{l1}, \ldots, x^0_{l,j-1},\, t,\, x^0_{l,j+1}, \ldots, x^0_{lp}\right)\right), \quad \text{for all } t \text{ in } (a_j,\, b_j),\ l = 1, \ldots, k$$

for the CPG(j), $j = 1, \ldots, p$.

Figure 2 shows k-CPG's of support vector machine (SVM) classifier for the subset of `iris` data. The number of clusters k is set to 4. One may see that the curves are more or less flat as functions of `sepal.length` and `sepal.width`, while the curves for `petal.length` and `petal.width` are clearly monotonically decreasing.

Even though k-CPG's are successful in summarizing n functions into k prototypes by k-means clustering, still they show the whole curves simply not taking into account the conditioning variates. Hence I devise one more procedure for crafting the curves.

Note that the batch of realized x_j values of which the observation unit belongs to specific cluster produces an plausible interval for the predictor. For instance, the k-means clustering of the `iris` subset data without the first variable `sepal.length` partitions the data set into groups of 22, 21, 31, 26 observations, among which `sepal.length` ranges over (4.9, 6.3), (6.2, 7.9), (5.4, 7.0), (5.6, 7.7). Similarly, one can obtain the plausible intervals for the other three variables by groups. See the left upper panel of Figure 3.

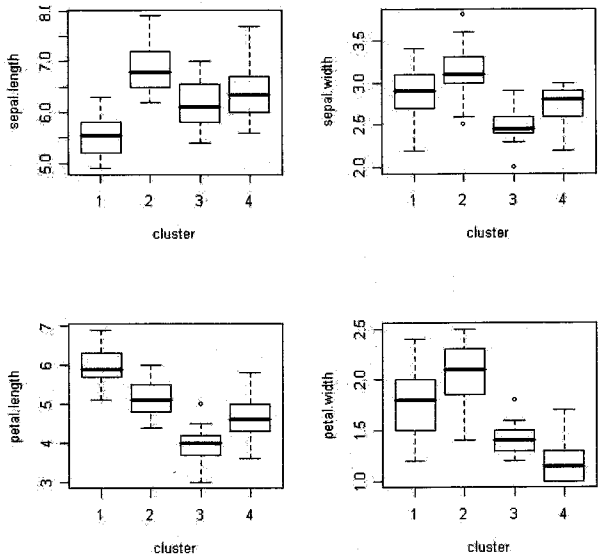


Figure 3: *Boxplots for four variables of* `iris` *subset data by k-means clusters*

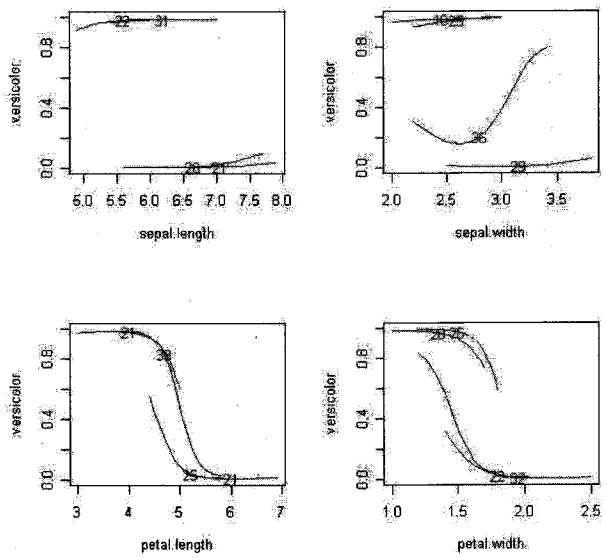


Figure 4: *Segmented k-CPG's of SVM classifier for* `iris` *subset data,* `Versicolor` *vs.* `Virginica`

Segmented k-CPG's: The Second Proposal

Draw the traces of the k-CPG's only over corresponding plausible intervals $(a_j^{(l)}, b_j^{(l)})$, $l = 1, \ldots, k$, for x_j $(j = 1, \ldots, p)$, where

$$a_j^{(l)} = \min_{i \in \text{cluster } l} x_{ij} \text{ and } b_j^{(l)} = \max_{i \in \text{cluster } l} x_{ij}.$$

Thus, the segmented k-CPG's of SVM classifier for the `iris` subset data will differ from the unsegmented k-CPG's of Figure 2 except supporting intervals. Figure 4 shows four panels of the

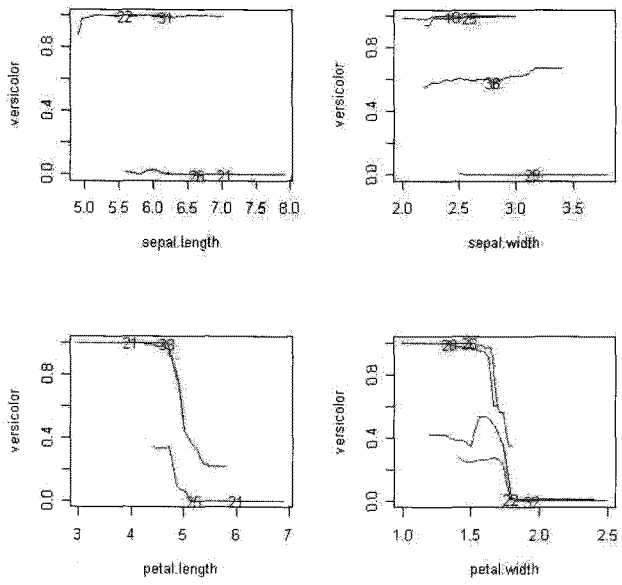


Figure 5: *Segmented k-CPG's of RF clasifier for* `iris` *subset data,* `Versicolor` *vs.* `Virginica`

segmented k-CPG's of SVM clasifier for the `iris` subset data. Figure 4 shows the same curves as those in Figure 2 on exposed portions.

One benefit of the segmented k-CPG's such as Figure 4 is the delivery of input variable importance measures and relative importance measures, that can be defined as

$$\text{Imp}(x_j) = \sum_{l=1}^{k} \frac{n_{l(j)}}{n} \left(\max_{a_j^{(l)} \le t \le b_j^{(l)}} f\left(x_{l1}^0, \ldots, t, \ldots, x_{lp}^0\right) - \min_{a_j^{(l)} \le t \le b_j^{(l)}} f\left(x_{l1}^0, \ldots, t, \ldots, x_{lp}^0\right) \right).$$

$$\text{Relative.Imp}(x_j) = \frac{\text{Imp}(x_j)}{\sum_{j=1}^{p} \text{Imp}(x_j)} \cdot 100(\%).$$

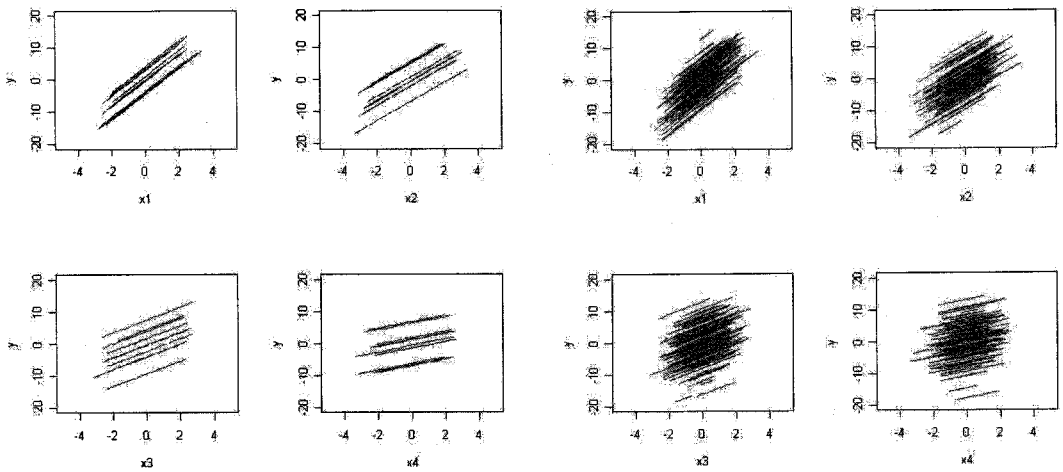
Relative importance measures of SVM clasifier for the `iris` subset data are 4.2%, 20.1%, 41.9%, 33.8% respectively for `sepal.length`, `sepal.width`, `petal.length`, `petal.width`. Hence we may conclude that `petal.length` and `petal.width` are two major predictors classifying `Versicolor` *vs.* `Virginica`.

So far, I have demonstrated the proposed method with a prediction function produced by a support vector machine (SVM). Of course, the generality of the method can be extended to any prediction function such as random forests (RF) model (Breiman, 2001). Figure 5 shows segmented k-CPG's of RF classifier for the `iris` subset data. Relative importance measures of RF classifier for the `iris` subset data are 3.9%, 6.1%, 44.6%, 45.4%. Figure 4 and Figure 5 share many common features.

3. Simulation Study

To see how the proposed methods work, a Monte-Carlo study is designed:

1) (X_1, X_2, X_3, X_4)'s are generated n times independently from a multivariate normal distribution



(a) *Case* $(\Sigma_0, 10)$ (b) *Case* $(\Sigma_0, 100)$

Figure 6: *Segmented k-CPG's for the simulation Cases* $(\Sigma_0, 10)$ *and* $(\Sigma_0, 100)$

with zero means and the covariance matrix Σ_0(and Σ_1).

$$\Sigma_0 = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.5 & 0.5 \\ 0.0 & 0.5 & 1.0 & 0.5 \\ 0.0 & 0.5 & 0.5 & 1.0 \end{pmatrix}.$$

2) For each (X_1, X_2, X_3, X_4), Y is generated according to

$$Y = 4X_1 + 3X_2 + 2X_1 + X_4 + \epsilon, \quad \epsilon \sim N\left(0, 0.5^2\right).$$

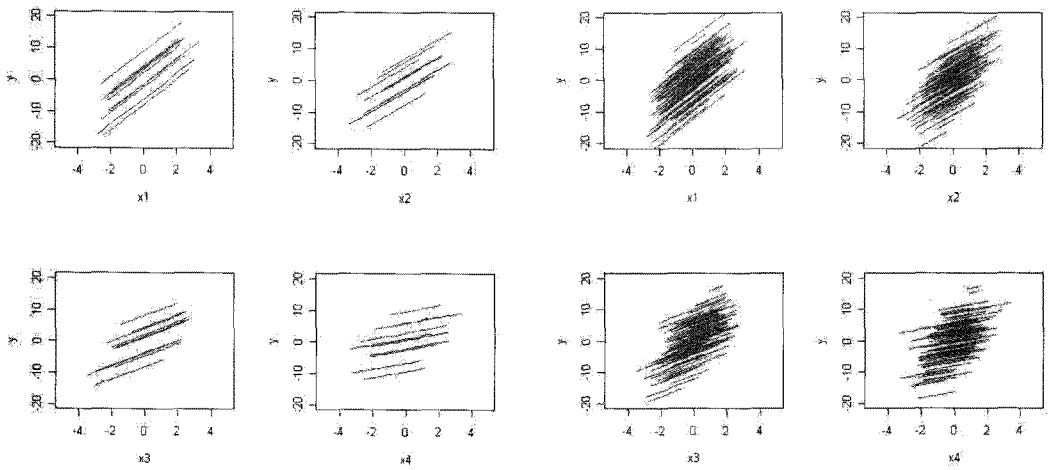
Number of observations n is set to 1,000. The number of clusters k is set to 10(and 100). The prediction model is obtained from linear regression. Each simulation case will be referred by Case (Σ, k).

Figure 6(a) and (b) are segmented k-CPG's, respectively, for Cases $(\Sigma_0, 10)$ and $(\Sigma_0, 100)$. All the graphs are right in the sense that the slopes of the panels are approximately 4, 3, 2, 1 as postulated. Also relative importance measures of predictors are approximately proportional to their coefficients of the underlying model:

$$\begin{aligned} \text{Case}\,(\Sigma_0,\ 10): &\ 38.9\%,\ 31.4\%,\ 19.6\%,\ 10.0\%. \\ \text{Case}\,(\Sigma_0, 100): &\ 40.7\%,\ 30.7\%,\ 18.7\%,\ 9.9\%. \end{aligned}$$

Figure 7(a) and (b) are segmented k-CPG's, respectively, for Cases $(\Sigma_1, 10)$ and $(\Sigma_1, 100)$. Four panels of the graphs show regression slopes near to 4, 3, 2, 1. Relative importance measures for input variables are, however, somewhat different. They turn out to be

$$\begin{aligned} \text{Case}\,(\Sigma_0,\ 10): &\ 43.8\%,\ 27.5\%,\ 19.6\%,\ 9.1\%. \\ \text{Case}\,(\Sigma_0, 100): &\ 45.9\%,\ 26.7\%,\ 18.7\%,\ 8.7\%. \end{aligned}$$



(a) *Case* $(\Sigma_1, 10)$ (b) *Case* $(\Sigma_1, 100)$

Figure 7: *Segmented k-CPG's for the simulation Cases* $(\Sigma_1, 10)$ *and* $(\Sigma_1, 100)$

But this disparity is understandable, since partial dispersions of each variable given the others are not equal for the cases of Σ_1. The partial standard deviations of X_1, X_2, X_3, X_4 given the others are 1.00, 0.82, 0.82, 0.82, respectively. Thus, theoretic relative importance for X_1 is

$$\frac{4 * 1.00}{4 * 1.00 + 3 * 0.82 + 2 * 0.82 + 2 * 0.82} = 44.9\%.$$

In that way, theoretic relative importances for X_2, X_3, X_4 are 27.5%, 18.4%, 9.2%. Hence, empirical results from simulated datasets are congruent to these theoretical numbers.

4. A Case of Environmental Study of Ozone

The case to be considered is the `ozone` data of Breiman and Friedman (1985). There are 330 observations of the response, ground level `ozone` (as a pollutant) in Los Angeles and nine explanatory variables: `vh`, the altitude at which the pressure is 500 millibars; `wind`, the wind speed(mph); `hum`, the humidity(%); `temp`, the temperature(F); `ibh`, the temperature inversion base height(feet); `dpg`, the pressure gradient(mmHg); `ibt`, the inversion base temperature(degrees F); `vis`, the visibility(miles); and `doy`, the day of the year.

Now, log(`ozone`) is modeled as a function of `temp`, `ibh`, `dpg`, `vis` and `doy` by support vector machine and random forests, of which the segmented k-CPG's are shown respectively in Figure 8 and Figure 9, with $k = 12$.

One can see that `temp` is the most important predictor in both fitted models. Such visual findings can be verified by numerical importance measures:

SVM: 32.2%, 14.9%, 18.2%, 13.6%, 21.0%.
RF: 36.5%, 19.0%, 15.8%, 8.2%, 20.6%.

In details, however, there are several differences between two regression models: `doy` and `dpg` are secondly important in SVM, while `doy` is in the second and `ibh` is in the third place in importance. Also, functional patterns of `doy` are different: In SVM the log(`ozone`) is peaked around `doy`

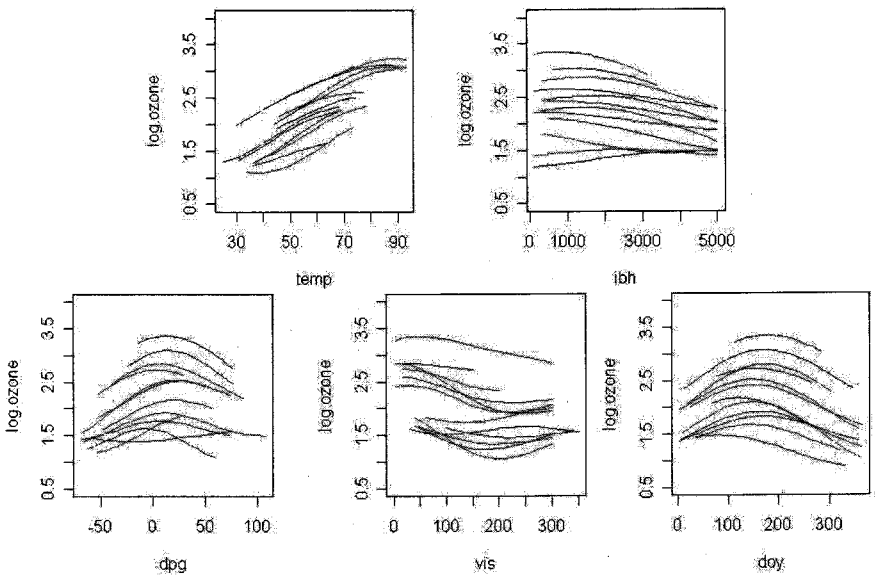


Figure 8: *Segmented k-CPG's of SVM regression for LA ozone data*

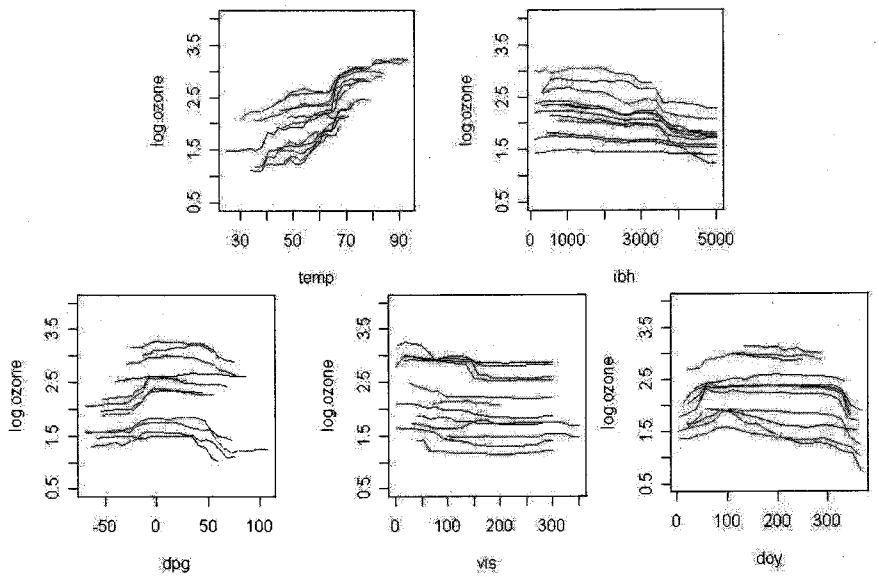


Figure 9: *Segmented k-CPG's of RF regression for LA ozone data*

= 150~180, while log(ozone) is almost flat for doy between 100 and 300. Breiman and Friedman (1985) discovered basically similar things, but they claimed the peak is at doy = 120, one or two months earlier than the finding of this study.

One may wonder how the graphs and measures change with different choice of k, the number of clusters. As experiment, set $k = 6$ and I obtained the very similar graphs (which are not shown here)

and measures as follows:

SVM: 32.1%, 14.5%, 19.4%, 13.0%, 20.9%.
RF: 36.2%, 22.0%, 14.3%, 8.2%, 19.2%.

Furthermore, I tried $k = 3$ and obtained the very similar results as follows:

SVM: 32.4%, 13.7%, 19.2%, 12.7%, 22.0%.
RF: 36.8%, 21.8%, 14.3%, 8.2%, 18.9%.

5. Concluding Remarks

In k-means clustering, it is well known that the determination of k, number of clusters, is an intricate task for data analysts. For the purpose of visualizing multi-variable functions, however, I think it is not that serious matter as verified in simulated settings of Section 3 and demonstrated in a real data case of Section 4. One may simply try various k's and observe the CPG's.

In the random forests (RF), variable importance measures are readily available which is defined by the magnitude of decrease in accuracy by randomly permuting data values of respective predictor (Breiman, 2001). Thus, RF's importance measure addresses different concept and does not take into account the effect of the other variables. Recently, there appeared an improvement (Strobl *et al.*, 2008). Compared to Breiman's measure of variable importance, the proposed method using segmented k-CPG's addresses directly the fitted prediction function.

References

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, **80**, 580–598.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

Huh, M. H. and Lee, Y. (2008). Simple graphs for complex prediction functions, *Communications of the Korean Statistical Society*, **15**, 343–351.

Strobl, C., Boulesteix, A., Kneib., T., Augustin, T. and Zeileis, A. (2008). Conditioning variable importance for random forests, *BMC Bioinformatics*, **9**, 307.

Received November 2008; Accepted November 2008