# Functional Data Analysis: A New Approach for Predicting Market Penetration of New Products

**Ashish Sood, Gareth M. James & Gerard J. Tellis**



Ashish Sood is Assistant Professor Marketing, Goizueta School of Business, Emory University, 1300 Clifton Rd NE, Atlanta, GA 30322; Tel: +1.404-727-4226, fax: +1.404-727-3552, E-mail: ashish_sood@bus.emory.edu.

Gareth James is Associate Professor of Statistics, Marshall School of Business, University of Southern California, P.O. Box 90089-0809, Los Angeles, California, USA; Tel: +1.213.740.9696, fax: +1.213.740.7313 E-mail: gareth@usc.edu

Gerard J. Tellis is Neely Chair of American Enterprise, Director of the Center for Global Innovation and Professor of Marketing at the Marshall School of Business, University of Southern California, P.O. Box 90089-1421, Los Angeles, California, USA. Tel: +1.213.740.5031, fax: +1.213.740.7828, E-mail: tellis@usc.edu

The study benefited from data from Emory Marketing Institute (www.emorymi.com), a grant from Goizueta Business School of Emory University, and a generous gift from Don Murray to USC Marshall's Center for Global Innovation.

# Functional Data Analysis: A New Approach for Predicting Market Penetration of New Products


## Abstract

The Bass (1969) model has been a standard for analyzing and predicting the market penetration of new products. The authors demonstrate the insights to be gained and predictive performance of Functional Data Analysis (FDA), a new class of non-parametric techniques that has shown impressive results within the statistics community, on the market penetration of 760 categories drawn from 21 products and 70 countries. The authors propose a new model called Functional Regression and compare its performance to over several models including the Classic Bass model, estimated means, last-observation projection, a meta-Bass model and an augmented meta-Bass model for predicting eight aspects of market penetration. Results a) validate the logic of FDA in integrating information across categories b) show that Augmented Functional Regression is superior to the above models and c) product specific effects are more important than country-specific effects when predicting penetration of an evolving new product.



Keywords: Predicting Market Penetration; Global Diffusion; Bass Model; Functional Data Analysis; Functional Principal Components; Generalized Additive Models; Functional Clustering; Spline Regression; New Products

## Introduction

Firms are introducing new products at an increasingly rapid rate. At the same time, the globalization of markets has increased the speed at which new products diffuse across countries, mature, and die off (Chandrasekaran and Tellis 2008). These two forces have increased the importance of the accurate prediction of the market penetration of an evolving new product. While research on modeling sales of new products in marketing has been quite insightful (Chandrasekaran and Tellis 2007; Peres, Mueller and Mahajan 2008), it is limited in a few respects. First, most studies rely primarily, if not exclusively, on the Bass model. Second, prior research, especially those based on the Bass model, need data past the peak sales or penetration for stable estimates and meaningful predictions. Third, prior research has not indicated how the wealth of accumulated penetration histories across countries and categories can be best integrated for good prediction of penetration of an evolving new product. For example, a vital unanswered question is whether a new product's penetration can be best predicted from past penetration of a) similar products in the same country, b) the same product in similar countries, c) the same product itself in the same country, or d) some combination of these three histories.

The current study attempts to address these limitations. In particular, it makes four contributions to the literature. First, we illustrate the potential advantages of using Functional Data Analysis (FDA) techniques for the analysis of penetration curves (Ramsay and Silverman, 2005). Second, we demonstrate how information about the historical evolution of new products in other categories and countries can be integrated to predict the evolution of penetration of a new product. Third, we compare the predictive performance of the Bass model versus an FDA approach, and some naïve models. Fourth, we indicate whether information about prior countries, other categories, the target product itself, or a combination of all three is most important in predicting the penetration of an evolving new product.

The model developed in this paper enables managers to solve the perennial problem in organizations of predicting penetration of new products in new markets. First, it provides managers a method to integrate information across similar products including competitors to achieve a superior prediction for an evolving new product. Second, the computationally efficient algorithms enable integration of information from a) past penetration of that category, b) past penetration of other categories, and c) knowledge of the product to which it belongs, to be used for prediction. The analysis can be implemented using standard statistical software and new data can be easily added to the analysis by managers. Third, the Augmented Functional Regression approach provides distinctly superior predictions to those from more standard models.

One important aspect of the current study is that it uses data about market penetration from most of 21 products across 70 countries, for a total of 760 categories (product x country combinations). The data include both developed and developing countries from Europe, Asia, Africa, Australasia, and North and South America. In scope, this study exceeds the sample used in prior studies (see Table 1). Yet the approach achieves our goals in a computationally efficient and substantively instructive manner.

Another important aspect of the study is that it uses Functional Data Analysis to analyze these data. Over the last decade FDA has become a very important emerging field in statistics, although it is not well known in the marketing literature. FDA provides a set of techniques that can improve the prediction of future items of interest especially in cases where prior longitudinal data is available for the same products, data is available from histories of similar products, or complete data is not available for some years. The central paradigm of FDA is to treat each function or curve as the unit of observation. We apply the FDA approach by treating the yearly cumulative penetration data of each category as 760 curves or functions. By taking this approach we can extend several standard statistical methods for use on the curves themselves.

For instance, we use functional principal components analysis (PCA) to identify the patterns of shapes in the penetration curves. Doing so enables a meaningful understanding of the variations among the curves. An additional benefit of the principal component analysis is that it provides a parsimonious, finite dimensional representation for each curve. In turn this allows us to perform functional regression by treating the functional principal component scores as the independent variables and future characteristics of the curves, such as future penetration or time to takeoff, as the dependent variable. We show that this approach to prediction is more accurate than the traditional approach of using information from only one curve. It also provides a deeper understanding of the evolutions of the penetration curves.

Finally, we perform functional clustering by grouping the curves into clusters with similar patterns of evolution in penetration. The groups that we form show strong clustering among certain products and provide further insights into the patterns of evolution in penetration. In particular plotting the principal component scores allows us to visually assess the level of clustering among different products for all 760 curves simultaneously. Such a visual representation would be impossible using the original curves.

The rest of the paper is organized as follows: The next three sections present the method, data and results. The last section discusses the limitations and implications of the research. Appendix A provides details of modeling of individual curves using splines. Appendix B provides details of k-means Clustering. Appendix C provides a glossary of technical terms used in the paper.

# Method

We present the method in five sections. The first three sections outline various applications of functional data analysis. Figure 1 provides a flowchart of the implementation of our three FDA techniques. The first section describes functional principal components. The second section shows how the functional principal component scores can be used to perform functional regression for predictions. The third section illustrates how the PCA scores can be used to perform functional cluster analysis and hence identify groupings among curves. The fourth section describes the alternate models against which we test the predictive performance of the FDA models. The last section details the method used for carrying out predictions.

## *Functional Principal Components*

Functional data analysis is a collection of techniques in statistics for the analysis of curves or functions. Most FDA techniques assume that the curves have been observed at all time points but in practice this is rarely the case. In some instances, curves may not be observed over all time periods. In other cases, the curves may only be observed over discrete intervals (e.g. annual estimates of adoption of new products). Since we have many observations for each curve we first use a simple smoothing spline approach to generate a continuous smooth curve from our discrete, observations. For example, a smoothing spline can be fit to a curve plotting the penetration of CD Players, given 10 years of discrete data to obtain its penetration curve. The full details of our spline implementation are provided in Appendix A.

We denote by $X_1(t)$, $X_2(t)$, …, $X_n(t)$ the n smooth curves that are our approximations to the penetration curves for each product-country combination and decompose these curves in the form,

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} e_{ij}\varphi_j(t) \qquad i = 1,\ldots,n \qquad \ldots(1)$$

subject to the following orthogonality constraints

$$\int \varphi_j^2(s)ds = 1 \quad \text{and} \quad \int \varphi_j(s)\varphi_k(s)ds \qquad \text{for} \qquad j \neq k .$$

The $\varphi_j(t)$'s represent the principal component functions, the $e_{ij}$'s the principal component scores corresponding to the $i^{th}$ curve and $\mu(t)$ the average curve over the entire population. As with standard principal components, $\varphi_1(t)$ represents the direction of greatest variability in the curves about their mean. $\varphi_2(t)$ represents the direction with next greatest variability subject to an orthogonality constraint with $\varphi_1(t)$ etc. The $e_{ij}$'s represent the amount that $X_i(t)$ varies in the direction defined by $\varphi_j(t)$. Hence a score of zero indicates that the shape of $X_i(t)$ is not similar to $\varphi_j(t)$ while a large score suggests that a high fraction of $X_i(t)$'s shape is generated from $\varphi_j(t)$.

To compute the functional principal components we divide the time period t=1 to t=T into p equally spaced points and evaluate $X_i(t)$ at each of these time points. Note that the new time points are not restricted to be yearly observations because the smoothing spline estimate can be evaluated at any point in time. Finally, we perform standard PCA on this p dimensional data. The resulting principal component vectors provide accurate approximations to the $\varphi_j(t)$'s at each of the p grid points and likewise the principal component scores represent the $e_{ij}$'s. We opted to set p=T and to evaluate the $\varphi_j(t)$'s at the original yearly time points. Since our penetration curves were generally smooth this approach generated smooth estimates for the $\varphi_j(t)$'s.

In theory, n different principal component curves are needed to perfectly represent all n $X_i(t)$'s. However, in practice a small number (D) of components usually explain a substantial proportion of the variability (Ramsay and Silverman, 2005) which indicates that

$$X_i(t) \approx \mu(t) + e_{i1}\varphi_1(t) + e_{i2}\varphi_2(t) + \cdots + e_{iD}\varphi_D(t) \qquad i = 1,\ldots,n \ldots(2)$$

for some positive D << n.

Note that the smooth functions, $X_i(t)$, are infinite dimensional in nature even though they are observed at only a finite number of time points. However, we use the $e_{ij}$'s in Equation (2) to reduce the infinite dimensional functional data to a small set of dimensions. This reduction in dimensions is crucial because it allows us to perform functional clustering and functional regression as described in the following two sections. In addition, it provides a parsimonious representation because it reduces the number of observations for each curve from T down to some small value D.

Note that even though the spline approach will not work in situations where only one or two time points are available for each curve, we can compute the functional principal components from sparsely observed data using other more sophisticated methods (see for example, James et al. 2000; Jank and Shmueli, 2006; Rettinger, Jank, Tutz, and Shmueli 2007). Hence the methods of functional clustering and functional regression that we describe in this paper can be applied even to products with only one or two years of penetration data.

### *Functional Regression*

We use functional regression to predict several items of interest, such as future marginal penetration level in any given year or the year of takeoff. Let $X_i$ (t) be the smooth spline representation of the i$^{th}$ curve observed over time such as the first five years of cumulative penetration for a given category. Let $Y_i$ represent a related item to be predicted, such as the marginal penetration in year six.

Functional regression establishes a relationship between predictor, $X_i(t)$, and the item to be predicted, $Y_i$, as follows:

$$Y_i = f(X_i(t)) + \varepsilon_i \qquad \mathrm{i} = 1,\ldots,\mathrm{n}. \qquad \ldots (3)$$

Equation (3) is difficult to work with directly because $X_i(t)$ is infinite dimensional. However, for any function f there exists a corresponding function g such that $f(X(t)) = g(e_1, e_2, ...)$ where $e_1$, $e_2$,… are the principal component scores of X(t). We use this equivalence to perform functional regression with the functional principal component scores as the independent variables. This approach is related to principal components regression which is often used for non-functional, but high dimensional, data. The simplest choice for g would be a linear function in which case Equation (3) becomes

$$Y_i = \beta_0 + \sum_{j=1}^{D} e_{ij}\beta_j + \varepsilon_j \qquad \ldots (4)$$

for some D≥ 1. A somewhat more powerful model is produced by assuming that g is an additive but non-linear, function (Hastie and Tibshirani, 1990). In this case, Equation (3) becomes

$$Y_i = \beta_0 + \sum_{j=1}^{D} g_j(e_{ij}) + \varepsilon_j \qquad \ldots (5)$$

where the $g_j$'s are non-linear functions that are estimated as part of the fitting procedure. There are different ways to model the $g_j$'s but one common approach, which we use in this paper, is the smoothing spline discussed in Appendix A. One advantage of using Equations (4) or (5) to implement a functional regression is that once the $e_{ij}$'s have been computed via the functional PCA, we can then use standard linear or additive regression to relate $Y_i$ to the principal component scores. We can also extend Equation (5) by adding covariates that contain information about the curves beyond the principal components, such as product or country characteristics or marketing variables.

## *Functional Clustering*

We use functional clustering for the purpose of better understanding the penetration patterns in the data. In particular, we wish to identify groups of similar curves and relate them to observed characteristics of these curves such as the product and country. We use the principal components described in the previous section to reduce the potentially large number of dimensions of variability and cluster all the curves in the sample.

We apply the standard k-means clustering approach (MacQueen 1967) to the D-dimensional principal component scores, $e_i$, described in Equation (2) to cluster all the curves in the sample. Appendix B provides more details of k-means clustering.

We use the "jump" approach (Sugar and James 2003) to select the optimal number of clusters, k. We compute $\xi_k = \gamma_k^{-Z} - \gamma^{-Z}_{k-1}$ for a range of values of k where $\gamma_k$ is given by (16) and Z is usually taken to be D/2. Sugar and James (2003) show through the use of information theory and simulations that setting the number of clusters equal to the value corresponding to the largest $\xi_k$ provides an accurate estimate of the true number of clusters in the data.

Once we compute the cluster centers, we assign each curve to its closest cluster mean curve. We can then use Equation (2) to project the centers back into the original curve space and examine the shape of a typical curve from each cluster.

## *Comparing Alternative Models*

To fully understand the advantages of FDA, we compare two implementations or models of FDA with five non-Functional models. We name the two functional models Functional Regression and Augmented Functional Regression and the five non functional models - Estimated Mean, Last Observation Projection, Classic Bass, Meta Bass, and Augmented Meta Bass. Table 2 classifies all the models based on their use of information across curves and nature of the model.

The functional regression approach has three main strengths. First it is able to incorporate information from other products to improve prediction accuracy. Second, it implements a non-parametric fitting procedure so it is not restricted by parametric assumptions. Third, it utilizes the functional nature of the penetration curves. We chose the five comparison models to gain an understanding of the gains from each of these strengths. For example, Classic Bass is parametric, does not use information from other products and is non-functional so it provides a baseline case where none of the strengths are present. The Meta Bass and Augmented Meta Bass models extend Classic Bass to incorporate information from other products but are still parametric and non-functional so they illustrate the improvement from borrowing strength across curves. The Last Observation Projection model uses information from all products and is also non-parametric so it illustrates the improvement from the first two strengths of FDA.

## Estimated Mean

The Estimated Mean is a simple model, which fits the mean of the item to be predicted in the estimation sample, as the predicted value of the item in the holdout sample. So, for example, to predict marginal penetration in year T+1 we use the mean marginal penetration in year T+1 among all curves in the estimation sample. Specifically, the prediction for the $i^{th}$ observation in the holdout sample, $\hat{Y}_i$, is given by

$$\hat{Y}_i = \overline{Y} \qquad \ldots (6)$$

where $\overline{Y}$ is the mean across all countries and products on the estimation sample. Note that this is a very simple model which does not use any information from the first T periods of data.

## Last Observation Projection

The Last Observation Projection is another simple model, which estimates the item to be predicted from only the last observation in each penetration curve. To do so, we first relate the item to be predicted, $Y_i$, to the final observed penetration level, $X_i(T)$, in the estimation sample. To estimate this relationship, we explore both a standard linear model (Equation 7) as well as a more flexible non-linear model (Equation 8),

$$Y_i = \beta_0 + \beta_1 X_i(T) + \varepsilon_i , \qquad \ldots (7)$$

$$Y_i = \beta_0 + g(X_i(T)) + \varepsilon_i \qquad \ldots (8)$$

We use the non-linear model for our final results. For the prediction, we use the estimated g from Equation (8) and the final observed penetration level ($X_i(T)$) in the holdout sample to get the predicted item in the holdout sample.

Note, this is a slightly superior model to the Estimated Mean, because it uses at least the last observation from each curve to be predicted. However, it still does not use any other prior data from the curve. We also tested out a linear regression model incorporating all T time periods, $X_i(1),\ldots, X_i(T)$, as independent variables. We have not reported the results here because, while this approach worked slightly better on some items and slightly worse on others, the overall results were not substantively different from the Last Observation predictions.

## Classic Bass

The Classic Bass Model (Bass 1969) fits each curve in the sample separately by estimating the following model:

$$s(t) = m\left[F(t) - F(t-1)\right] + \varepsilon(t), \qquad F(t) = \frac{1 - e^{-(p+q)t}}{1 + (q/p)e^{-(p+q)t}} \qquad \ldots (9)$$

where t=time period, s(t)=marginal penetration at time t, p=coefficient of innovation, q=coefficient of imitation and m=final cumulative penetration.

We estimate the model via the genetic algorithm because Venkatesan et al (2004) provide convincing evidence that the genetic algorithm provides the best method for fitting the Bass model relative to all prior estimation methods. For each curve, we use the first T years of data to estimate the three Bass parameters, m, p and q. We then predict the next five years of penetration levels by plugging the estimated parameters back into the Bass model and evaluating at times T+1 through T+5. We predict the time of peak marginal penetration by using t=log(q/p)/(p+q) and the peak marginal penetration using s= m(p+q)2/4q. We do not predict time to takeoff with the Classic Bass Model. Note that the Classic Bass Model does not distinguish between holdout and estimation samples because each curve is fit individually without using information from other curves.

### Meta-Bass

In the Meta-Bass model, we extend the Classic Bass Model to use information across curves. To do so, we first estimate m, p and q for each curve using the genetic algorithm, as outlined above. Then, for each item to be predicted, we fit the non-linear additive model,

$$Y_i = \beta_0 + g_1(m_i) + g_2(p_i) + g_3(q_i) + \varepsilon_i \ , \qquad \ldots (10)$$

to the estimation sample where $g_1$, $g_2$, and $g_3$ are smoothing splines as defined previously. We use the estimated parameters from this additive model and the estimates of m, p, and q for each curve in the holdout sample to compute the corresponding item to be predicted for each of the holdout curves. Note that the estimation of m, p, and q can also be done using a Bayesian formulation with a prior on {m,p,q}.

### Augmented Meta-Bass

The Augmented Meta-Bass is the same non-linear additive model used for the Meta-Bass except that we add an indicator variable for each of the R products to which each curve belongs, thus:

$$Y_i = \beta_0 + g_1(m_i) + g_2(p_i) + g_3(q_i) + \sum_{r=1}^{R-1} \delta_r I_{ir} + \varepsilon_i \qquad \ldots (11)$$

where $I_{ir}$ =1 if the ith curve belongs to product r and 0 otherwise and the $\delta_r$'s are regression coefficients that are estimated as part of the model fitting procedure. Note, that the Meta-Bass and Augmented Meta-Bass are extensions of the Classic Bass that make use of all of the information across curves, rather than just utilizing each curve individually. Since, using information across curves is an essential feature of functional regression, doing so puts the Meta Bass and the Augmented Bass on the same platform as the FDA models (see Table 2).

## Functional Regression

For the Functional Regression model, we compute four principal component scores, the first two each on the penetration curves, $X_i(t)$, and on the velocity curves, $X_i'(t)$. The principal component scores on the velocity curves are computed in an identical fashion to that for the penetration curves except that we utilize the derivative of $X_i(t)$. We then use these four scores as the independent variables in an additive regression model, as shown in Equation (8), on the estimation sample. We then use the estimated parameters of this equation and the data from the curves in the holdout sample, to compute the items to be predicted in the holdout sample.

## Augmented Functional Regression

Our second functional approach enhances the power of Functional Regression by adding an indicator variable for each of the R products to which each curve belongs, as with the Augmented Meta-Bass model. Hence, the Augmented Functional Regression model involves estimating a non-linear additive model on the estimation sample as follows

$$Y_i = \beta_0 + \sum_{j=1}^{4} g_j(e_{ij}) + \sum_{r=1}^{R-1} \delta_r I_{tr} + \varepsilon_i \qquad \ldots (12)$$

where the $g_j$'s are smoothing splines. We then compute the items to be predicted for each curve in the holdout sample from the estimated values of the above parameters and the data in each curve in the holdout sample. This model is directly comparable to Augmented Meta Bass as both models use information across curves and from products.

## *Method for Prediction*

We explain the specific procedure for carrying out the prediction in three parts: items being predicted, computation of errors, and partitioning of sample.

### Items Being Predicted

We first truncate each curve at the $T^{th}$ year. We use the penetration in years 1 to T to estimate the model and predict the marginal change in penetration for years T+1 to T+5. For each curve, we also predict the number of years to takeoff, the years to peak marginal penetration, and the level of peak marginal penetration. Takeoff is the first turning point in sales, marking the transition from the introductory to the growth stage of the product life cycle. We identify the year of takeoff based on the definition proposed by Golder and Tellis (1997). Thus, we predict a total of eight items for each of seven models, for a total of 56 model-items. We do this whole process once each for T=5 years and T=10 years.

### Computation of Errors

For each of these 56 model-items to be predicted, we compute the mean absolute deviation (MAD) over all penetration curves, i.e.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \qquad \dots (13)$$

where $Y_i$ is a particular item for curve i and $\hat{Y}_i$ is the corresponding estimate using a given model.

**Partitioning of Sample**

We use ten-fold cross-validation by randomly partitioning the curves into ten equal groups. We hold out one group, estimate each of the models on the remaining nine groups using data from years 1 to T and then form predictions on the held out group for years T+1 to T+5. We repeat this process ten times, for each of the ten held-out groups of data. T is the same for all countries and products. Figure 2 provides a graphical description of our process. For each of the 56 model-items, we compute the mean absolute deviation as an average of these ten iterations. Note, that k-fold cross-validation is superior to simple splitting of data into one holdout and training group, because all of the data are used (randomly) as a holdout once.

# Data

This section details our sample, sources, and procedure for data collection.

## *Sample*

Most of the prior studies are limited in scope in terms of both product type and geographical breadth (see Table 1). We collect data on 760 categories drawn from 21 products (see Table 3) and 70 countries (see Figure 3). The sample includes a broad sample from three categories - household white goods, computers and communication, and entertainment and lifestyle.

## *Sources*

The information required for this study is penetration rates of different products introduced in different markets from the year of introduction to at least some time after the takeoff. The primary source of our data is the Global Market Information Database of Euromonitor International. Euromonitor International’s Global Market Information Database is an integrated online information system that provides business intelligence on countries, consumers and lifestyles. We also use press releases, industry reports and archived records to identify the year of introduction from databases like Factiva and Productscan.

### *Procedure*

We follow the general rules for data collection for the historical method (Golder 2000). We explain specific problems we encounter and the rules we use to resolve them. We screen the categories to be used by three criteria. First, we suspect that all curves that have penetration rates above 1% in the first year, may have missing early years of data. So, for these categories, we check the year of introduction from historical reports or press releases. We exclude all categories where data is not available from the first year of introduction. Second, we exclude from our analysis any categories that do not contain at least T+5 years of observations or have not reached peak marginal penetration. Third, the data from this source is only available from 1977. Hence, we exclude all categories where the product had been introduced or taken off earlier than 1977.

## Results

We present the results on functional principal components, functional regression, and functional clustering.

### *Functional Principal Components*

Figures 4a and 4b provide plots of $\varphi_1(t)$ and $\varphi_2(t)$ computed from the first ten years of observations on the 760 penetration curves. The first principal component represents the amount by which a curve's penetration, at year ten, is above or below the global year ten average of all 760 curves. Categories with a positive score on the first component end up with above average last period penetration levels while those with negative scores have below average last period penetration. Alternatively, the second principal component represents the way that the penetration levels evolve. Categories with a positive score on the second component grow most rapidly in the early years but slow down by year ten while those with a negative score are associated with slow initial growth and a rapid increase towards year ten.

An alternative way of visualizing these curves is presented in Figures 4c and 4d. Here the black line corresponds to $\mu(t)$, the average penetration level over all 760 curves. The red lines represent $\mu(t) \pm \eta_j \varphi_j(t)$ where $\eta_j$ is a constant proportional to the standard deviation of $e_{ij}$. Figure 4c shows that categories with a positive value for $e_{i1}$ will have above average last period penetration levels at year ten while ones with a negative $e_{i1}$ will remain stagnant over time and will have last period penetration levels below the overall average. Alternatively, Figure 4d shows that categories with a positive value for $e_{i2}$ will grow somewhat faster than average to begin with but then fall below average after 10 years while curves with a negative $e_{i2}$ will have the opposite pattern.

Remarkably, $\varphi_1(t)$ and $\varphi_2(t)$ together explain over 99% of the variability in the smoothed penetration curves which indicates that $e_{i1}$ and $e_{i2}$ provide a highly accurate two-dimensional representation of $X_i(t)$. However, it should be noted that the smoothing spline approach removes some of the variability in the data so $\varphi_1(t)$ and $\varphi_2(t)$ explain somewhat less than 99% of the variation in the observed penetration data. As mentioned previously one can also compute principal components for the velocity curves of the penetration levels. When we perform this decomposition on the penetration curves the principal components of $X'_i(t)$ have a very similar structure to those for $X_i(t)$.

## Functional Regression

This section presents the performance of the seven models on the eight items to be predicted. Tables 4a and 4b present the cross-validated mean absolute deviation scores for each model using cutoffs of T=5 and T=10 years of training data respectively. We also compute the fraction of curves for which Augmented Functional Regression outperforms each of the other methods (see Tables 5a and 5b).

In order to assess the ability of functional data analysis to predict items of penetration curves, we compare the Functional Regression model to the five non-functional models. Functional Regression is superior to Estimation Mean, Last Observation Projection, and Classic Bass at predicting all eight items at both cutoff times (see Table 4). The reason is that the Estimation Mean and the Last Observation Projection use minimal information from prior time periods while Classic Bass uses no information across curves. Functional Regression is also better than Meta Bass on all items for both cutoff times except for time to peak marginal penetration at cutoff time T=10 years.

The performance of Functional Regression is mixed when compared with Augmented Meta Bass. At the cutoff of T=5 years, Functional Regression is superior to Augmented Meta-Bass for the T+1, T+2, and T+3 years, similar for T+4 years but inferior for the other four items (see Table 4a and Table 5a). At the cutoff of T=10 years, Functional Regression outperforms Augmented Meta-Bass for T+1 through T+5 years as well as time to takeoff but not for time to peak marginal penetration and peak marginal penetration (see Table 4b and Table 5b). The reason is that the Augmented Meta Bass uses information about product while the Functional Regression does not.

On the other hand, with the sole exception of the Classical Bass predicting year T+1 with cutoff of T=10, the Augmented Functional Regression model is superior to all non functional models including Augmented Meta Bass, for every item to be predicted and for both cutoff times. The Augmented Functional Regression is also superior to Functional Regression, except in three instances where it is equal or slightly inferior (for T+1 years at cutoff of T=5 years and

T+1, T+4 years at cutoff of T=10 years). The superiority over Functional Regression is most noticeable in the time to takeoff and time to peak marginal penetration.

When considering Table 5, Augmented Functional Regression is superior for at least 50% of the curves in 94 out of the 96 possible comparisons with other models. It appears that the Functional Regression model is slightly superior for predicting T+1 but the augmented version is preferable for any longer range predictions. Most of the differences in Tables 4 and 5 are highly statistically significant. We also tested out the Augmented Functional Regression model with the addition of a predictor for geographic region as defined in the clustering section but found that the performance deteriorated slightly. In summary, the Augmented Functional Regression model outperforms other models in over 96 % of the comparisons with six alternate models to predict seven items across two cutoff times.

## *Functional Clustering*

Figure 5 provides several approaches to viewing the results from the functional clustering using k-means on $e_{i1}$ and $e_{i2}$. The "jump" approach of Sugar and James (2003) suggests between six and nine clusters. We opt for six to provide the most parsimonious representation (see Table 6). Figure 5a plots the centers of the six clusters on the original time domain. The figure illustrates the pattern of growth of a typical curve in each cluster. Alternatively, Figure 5b plots all 760 curves in the reduced two dimensional space, using the same colors to represent each cluster as for Figure 5a. The six cluster centers are represented as solid black circles.

Each cluster differs from the other clusters in the pattern of penetration over time. Broadly speaking, Clusters 1 through 3 represent high growth products while the last three correspond to lower growth rates. Cluster 1 takes on large values in both the first and second principal component dimensions. Recall that a positive value in the first dimension corresponds to overall high last period penetration while a positive value in the second dimension represents a fast growth at the beginning but a slow down by year ten. The black curve in Figure 5a shows this pattern with the fastest overall growth but a slight slowdown by year ten. Cluster 2 is close to zero for the second dimension indicating no overall slowdown as we can see from the pink curve. Clusters 3 and 4 provide an interesting contrast. Cluster 3 has a negative value in the second dimension while Cluster 4 is positive. This suggests a slow start for Cluster 3 but with increasing momentum by year ten and the opposite pattern for Cluster 4. Figure 5a shows precisely this pattern with Cluster 4 starting ahead of Cluster 3 but then falling rapidly behind. Cluster 5 represents a moderate rate of growth while Cluster 6, which contains the largest number of products, corresponds to a much slower improvement in penetration.

We also examine whether the penetration patterns differ across products. Figure 5c illustrates the growth patterns for the twenty-one different products in the sample. We plot all 760 curves in our two-dimensional space using a different plotting symbol for each product. There are very clear patterns within the same product. For example, the green stars correspond to internet-compatible personal computers and have almost uniformly large values on the first dimension indicating rapid increases in penetration levels. Notice that one product may have both positive and negative values for the second dimension, suggesting more rapid takeoff in some markets over others. Alternatively, the yellow squares represent DVD players and have a very tight clustering with almost uniformly moderate scores on the first principal component and negative scores on the second principal component. These results suggest a slow initial growth with much more rapid expansion towards year ten. The tighter clustering suggests that the takeoff for these products is largely similar across different markets in the sample. Finally, the blue solid dots, representing Video Tape Recorders, show the opposite pattern with large positive scores on the second dimension suggesting fast initial growth but then a slow down in later years.

Curves for each product are from a variety of countries. Table 6 provides the fraction of curves of each product that fall within each of the six clusters. The functional clustering suggests three groups - fast growth electronics, slower growth electronics, and household goods. The first three clusters capture a group of six fast growth electronics products with Cluster 1 primarily internet personal computers, Cluster 2 a mixture and Cluster 3 mainly DVD players. The other three clusters capture a group of slow growth products: Video game consoles, satellite TV, and CD players make up the bulk of Cluster 4. Cluster 5 contains many products but seems to principally concentrate on countries with slower growth for CD and DVD players, Satellite TV, and Video game consoles. Finally, Cluster 6, the slowest growth cluster, contains the vast bulk of household appliances.

Similarly, we also examine whether the penetration patterns differ across countries. We categorize the data into seven economic groupings (see Table 7 and Figure 3). For each group, Table 7 shows the fraction of curves that fall in each of the six clusters. For example, for countries from Africa and developing Asia 86% of curves fall into the slowest growth Cluster 6. In contrast, North American, Western Europe and Australasia have curves that are more spread out over the six clusters, with only 30% in the slowest growth Cluster 6.

## Discussion

Predicting the market penetration of new products is currently growing in importance due to increasing globalization, rapid introduction of new products, and rapid obsolescence of newly introduced products. Moreover, good record keeping has generated a wealth of new product penetration histories. The Bass model has been the standard model for analyzing such histories.

However, the literature has not shown how exactly researchers should integrate the rich record of penetration histories across categories with the penetration of an evolving new product to predict future characteristics of its penetration. Functional data analysis, which has gained significant importance in statistics, is well suited for this task. Our goal is to demonstrate and assess the merit of functional data analysis for predicting the market penetration of new products and compare it with the Bass model.

We compare the predictive performance of Functional Regression and Augmented Functional Regression with five other models: two simple or naïve models, the Classic Bass model, the Meta Bass model and the Augmented Meta-Bass Model on eight items to be predicted.

Our analysis leads to three important results

1. The essential logic of integrating information across categories, which is the foundation of functional data analysis, provides superior prediction for an evolving new product.
2. Specifically, an evolving category can be best predicted by integrating information from a) past penetration of that category, b) past penetration of other categories, and c) knowledge of the product to which it belongs, via the framework of functional regression.
3. For a vast variety of items that need to be predicted, the Augmented Functional Regression is distinctly superior to a variety of models, including simple or naïve models classic and enhanced Bass models and Functional Regression.

## Implications

Our functional regression has at least two clear managerial implications. First, our method can be used to make more accurate predictions of the future trajectory for both existing products as well as new products with only a few years of observations. One could also make predictions for the evolution of a new product without any data based on the previously observed principal component scores of similar products. Second, although we have not done so here, it would be conceptually simple to add additional variables such as pricing and advertising information to the functional regression model. The addition of these variables would not only allow a manager to passively predict but to also control future penetration levels.

Our results raise the following questions with further managerial and research implications.

First, why don't simple extrapolative models work well for prediction, as some researchers assert they do (Fader and Hardie 2005; Armstrong 1984; 1978; Armstrong and Lusk 1983). Our analysis makes it clear that there are two dimensions of information that are not captured by simple models. One, there is valuable information in the prior history of the new product, which as the Bass model suggests, probably arises from consumers' innovative and imitative tendencies. Two, there is intrinsic information across products and countries, which may be used effectively to predict the penetration of an evolving new product. Despite their intuitive appeal, simple models that do not capture these sources of information will fail to predict well.

Second, why does the Classic Bass model not work as well for prediction? We suspect that it does not fully capture the two dimensions of information. One, the classic Bass model ignores other categories. This fact is borne out by the superiority of the Meta Bass and the Augmented Meta Bass in predicting items further out into the future. Both of these latter models capture information from other categories. Thus, even in a parametric setting, increased predictive accuracy can be gained by incorporating information from multiple categories, especially when predicting further into the future. Two, the classic Bass model is relatively flexible but nevertheless parametric. So it is limited in the range of shapes that it can take on. In particular, it is constrained to symmetric shapes for certain values of p and q. The relatively strong performance of the Last Observation Projection model shows that removing the parametric assumptions can cause additional improvements.  In line with that result, FDA provides higher flexibility by using a non-parametric approach. So it can capture a variety of flexible patterns without over fitting, with the help of the principal components as explained earlier. The main disadvantage of a non-parametric method is that the increased flexibility can produce variability in the estimates. However, functional regression builds strength across the 760 curves to mitigate the problem of variability while generating more flexible estimates than those produced by the Classic Bass model.

Third, why does the Augmented Functional Regression outperform Functional Regression, especially for items further out into the future? The probable reason is that a particular product has a distinct pattern of penetration over time. Adding knowledge of that product further stabilizes the variability of predictions around their true value. This pattern can be seen in both the improvement of Augmented Meta-Bass over Meta-Bass and Augmented Functional Regression over Functional Regression. Note also, that the improvement is greatest in peak marginal penetration, an item which arguably is most closely associated with a product.

Fourth, why is product seemingly more relevant for predicting market penetration than is country? The probable reason is that the evolution of market penetration seems to follow more distinct patterns by the nature of the product than by the country. For example, electronic products with universal appeal diffuse rapidly across countries both large and small and developed and developing. On the other hand, culturally sensitive products such as food appliances diffuse slowly overall and very differently across countries. Moreover, our data is only after 1977. Due to increasing industrialization of developing countries and flattening of the world economy, inter-country differences are much smaller after 1977, than before it.

Fifth, is the exclusion of marketing variables a limitation of Functional Regression? We posit that it is not. Indeed, we show the superiority of Augmented Functional Regression, which includes a covariate for the product to which the curve belongs. In like manner, this model could also include covariates for marketing variables, such as price, quality, or advertising.

To illustrate some of the above points, Figure 6 demonstrates six plots of the predictive performance of the Classic Bass model (red) and the Functional Regression model (green) relative to actual (black). These six plots are drawn from among those where Functional Regression does the best. For each plot, the first ten periods are fitted on the estimation sample, while the last five periods are predictions on the hold out sample. Both models do well in the estimation periods. However, performance varies dramatically in the hold out periods.

Note how for curves, 676, 557, and 126, a generally flat curve with a late takeoff in the last two periods, tricks the classic Bass into over predicting penetration for the holdout period. However, Functional Regression, which draws strength from other categories, is not so influenced by the last two periods. Also, note how for curves 582, 572, and 121, the parameterization of the Bass model leads it to predict symmetric curves which are quite far from the actual.

This study has the following limitations. First, while the data is from a single source, the source itself does not record data before 1977. Indeed, we drop categories in some countries where we consider the year of introduction precedes 1977. We also drop categories in countries where penetration is not high enough till 2006. So the data are not balanced by country. Thus, substantive estimates about time to takeoff or about penetration by countries must be made with caution. However, that fact should not affect the comparison of the models, because all models have access to the same data. Second, depending on the release patterns of a particular product the product predictor used in Augmented Meta-Bass and Augmented Functional Regression may or may not be available. Third, our data do not include any marketing variables. However, the strength of the Augmented Functional Regression is that it can include such marketing variables. Fourth, our approach applies to the prediction of market penetration using only aggregate historical data. Other approaches exist to predict based on survey and experimental data (Hauser, Tellis and Griffin 2006) and disaggregate historical data (Tellis and Franses 2006). Future research could address how better improvements can be obtained by using such information when available. Fifth, future research could also address the of functional regression for predicting the evolution of underlying technologies (e.g., Sood and Tellis 2005).

## Appendix A

## Modeling of Individual Curves

Suppose that a curve, X(t), has been measured at times t=1, 2,…,T. Then the smoothing spline estimate is defined as the function, h(t), that minimizes

$$\sum_{t=1}^{T}\left(X(t)-h(t)\right)^2+\lambda\int\{h''(s)\}^2\,ds \qquad (14)$$

for a given value of λ>0 (Hastie et al., 2001). The first squared error term in Equation (14) forces h(t) to provide a close fit to the observed data while the second integrated second derivative term penalizes curvature in h(t). The tuning parameter λ determines the relative importance of the two components in the fitting procedure. Large values of λ force a h(t) to be chosen such that the second derivative is close to zero. Hence as λ gets larger h(t) becomes closer to a straight line, which minimizes the second derivative at zero. Smaller values of λ place more emphasis on h(t)'s that minimize the squared error term and hence produce more flexible estimates. We follow the standard practice of choosing λ as the value that provides the smallest cross-validated residual sum of squared errors (Hastie et al., 2001). Remarkably, even though Equation (14) is minimized over all smooth functions it has been shown that its solution is uniquely given by a finite dimensional natural cubic spline (Green and Silverman, 1994), which allows the smoothing spline to be easily computed. A cubic spline is formed by dividing the time period into L regions where larger values of L generate a more flexible spline. Within the $l^{th}$ region a cubic polynomial of the form

$$h(t)=a_l+b_l t+c_l t^2+d_l t^3 \qquad (15)$$

is fit to the data. Different coefficients, $a_l$, $b_l$, $c_l$ and $d_l$ are used for each region subject to the constraints that h(t) must be continuous at the boundary points of the regions and also have continuous first and second derivatives. In a natural cubic spline, the second derivative of each polynomial is also set to zero at the endpoints of the time period. In the more complicated situation where the curves are sparsely observed over time (e.g. due to a different data generating process or data limitations), a number of alternatives have been proposed. For example, James et al (2000) suggest a random effects approach when computing sparsely observed curves.

# Appendix B

# K-means Clustering

k-means clustering works by locating D-dimensional cluster centers $c_1, \ldots, c_k$ which minimize the sum of squared distances between each observation and its closest cluster center i.e. find $c_1, \ldots, c_k$ to minimize

$$\gamma_k = \sum_{i=1}^{n} \min_{c_1, c_2, \ldots, c_k} \left\| e_i - c_j \right\|^2 \quad (16)$$

We use an iterative algorithm to minimize $\gamma$. First we choose an initial set of candidate centers, $c_1, \ldots, c_k$, by randomly selecting k of the $e_i$'s and assign each curve to its closest center. Then, for each cluster, we define a new center by taking the average over all curves currently assigned to that cluster. We continue this algorithm until additional iterations do not yield significant changes in the cluster centers.

# Appendix C

# Glossary

**Augmented Meta Bass**: The Augmented Meta-Bass Model extends the Meta Bass Model (see below) with the inclusion of an additional indicator for the product to which each curve belongs.

**Augmented Functional Regression**: Augmented Functional Regression extends Functional Regression (see below) with the inclusion of an additional indicator for the product to which each curve belongs. (see also Functional Regression)

**Category**: A category is a product-country combination, such as dishwasher in Spain.
.

**Classic Bass Model**: Is the original Bass model (Bass 1969), which models the current sales of a new product as a function of prior sales and cumulative sales. The model derives from the assumption that a consumer's adoption of a new product is a function of some adoption rate plus the number of prior adopters. We fit the model using a Genetic Algorithm.

**Estimated Mean**: The Estimated Mean model uses the mean of the item in the estimation sample, as the value of the item to be predicted in the holdout sample.

**Functional Clustering**: Functional clustering is a method of grouping functions with similar characteristics e.g. to find products with similar penetration histories.

**Functional Data Analysis**: Functional data analysis is a collection of statistical techniques for the analysis of curves or functions.

**Functional Principal Components**: Functional Principal Components is a statistical technique to identify the main ways that a group of curves tends to differ from the overall mean function. One application is to reduce infinite dimensional functions to a finite number of dimensions.

**Functional Regression**: Functional Regression is a model used to establish the relationship between a dependent variable and an independent variable. The key difference between a functional and a standard regression model is that in the latter case the independent variable is a function e.g. a smooth representation of a curve observed over time.

**Historical Method**: The historical method is a scientific method where researchers gather data from two or more published sources that meet criteria of reliability, independence, objectivity, corroboration, and contemporaneity.

**k-means Clustering**: k-means clustering is a method of partitioning data into different sets such that each point is assigned to the cluster whose center is nearest.

**Last Observation Projection**: The Last Observation Projection model estimates the item to be predicted from only the last observation in each penetration curve.

**Market Penetration**: Market Penetration is defined as the percentage of households in a geographical area that have adopted (purchased) a certain product.

**Meta Bass**: The Meta-Bass model is a regression model that relates estimates of m, p, and q to the item to be predicted for a set of curves.

**Product**: A good that meets a distinct consumer need (e.g. dishwasher)

**Smoothing Spline**: Smoothing spline is a curve that provides a smooth approximation to a set of points observed on an (X,Y) axis.

**Takeoff**: Takeoff is the first turning point in sales, marking the transition from the introductory to the growth stage of the product life cycle.

**Cross-validation**: Cross-validation involves partitioning a sample of data into subsets. A model is then fit using the data from all but one of the subsets and its accuracy is assessed on the remaining data set. This fitting and accuracy assessment procedure is then repeated for each subset. The overall accuracy of the model is estimated by averaging the results for each subset.

***Table 1:***

***Scope of Prior Studies****

| Authors | Categories | Countries |
|---|---|---|
| Gatignon, Eliashberg and Robertson (1989) | 6 consumer durables | 14 European countries |
| Mahajan, Muller and Bass (1990) | Numerous studies | |
| Sultan, Farley and Lehmann (1990) | 213 applications | US, European countries |
| Helsen, Jedidi and DeSarbo (1993) | 3 consumer durables | 11 European countries and US |
| Ganesh and Kumar (1996) | 1 industrial product | 10 European countries, US, Japan |
| Ganesh, Kumar, Subramaniam (1997) | 4 consumer durables | 16 European countries |
| Golder and Tellis (1997) | 31 consumer durables | USA |
| Putsis et al (1997) | 4 consumer durables | 10 European countries |
| Dekimpe, Parker and Sarvary (1998) | 1 service | 74 countries |
| Kumar, Ganesh and Echambadi (1998) | 5 consumer durables | 14 European countries |
| Golder and Tellis (1998) | 10 consumer durables | USA |
| Kohli, Lehmann and Pae (1999) | 32 appliances, house wares and electronics | USA |
| Dekimpe, Parker and Sarvary (2000) | 1 innovation | More than 160 countries |
| Mahajan, Muller and Wind (2000) | Numerous studies | |
| Van den Bulte (2000) | 31 consumer durables | USA |
| Talukdar, Sudhir, Ainslie (2002) | 6 consumer durables | 31 countries |
| Agarwal and Bayus (2002) | 30 innovations | USA |
| Goldenberg, Libai and Muller (2002) | 32 innovations | USA |
| Tellis, Stremersch and Yin (2003) | 10 consumer durables | 16 European countries |
| Golder and Tellis (2004) | 30 consumer durables | USA |
| Stremersch and Tellis (2004) | 10 consumer durables | 16 European countries |
| Van den Bulte and Stremersch (2004) | 293 applications | 28 countries |
| Chandrasekaran and Tellis (2007) | 16 products and services | 40 countries |

* Adapted from Chandrasekaran and Tellis (2008)

*Table 2:*

*Classification of Models*

| | | Analysis of Curves | |
|---|---|---|---|
| | | **Non Functional Analysis** | **Functional Analysis** |
| **Uses Information Across Curves** | **No** | Classical Bass | |
| | **Yes** | Estimation Mean<br>Last Observation Projection<br>Meta-Bass<br>Augmented Meta-Bass | Functional Regression<br>Augmented Functional Regression |

***Table 3: Sample Categories***

| Entertainment and Lifestyle | Household White Goods | Computers/ Communication |
|---|---|---|
| Cable TV | Air conditioner | Internet PC |
| Camera | Dishwasher | Personal Computer |
| CD Player | Freezer | Fax |
| Color TV | Microwave Owen | Satellite TV |
| DVD Player | Tumble Drier | Telephone |
| Hi-Fi Stereo | Vacuum Cleaner | |
| Video Camera | Washing Machine | |
| Video Tape Recorder | | |
| Videogame Console | | |

***Table 4:***

***Mean Absolute Deviation by Model and Item***

***Table 4a: Using five years (T=5) of training data***

| **Item to be Predicted** | **Method** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Estimation Mean** | **Last Observation Projection** | **Classical Bass** | **Meta-Bass** | **Augmented Meta-Bass** | **Functional Regression** | **Augmented Functional Regression** |
| T +1 | 9.08 | 4.05 | 3.01 | 7.47 | 7.52 | 2.43 | 2.48 |
| T +2 | 12.39 | 7.49 | 7.18 | 10.50 | 10.01 | 5.46 | 5.12 |
| T +3 | 14.49 | 10.01 | 12.40 | 12.74 | 11.20 | 8.14 | 6.87 |
| T +4 | 17.35 | 13.74 | 17.27 | 16.08 | 12.16 | 11.75 | 8.29 |
| T +5 | 19.57 | 17.28 | 19.52 | 19.52 | 13.99 | 15.82 | 9.85 |
| Takeoff | 3.36 | 2.84 | NA | 2.69 | 2.41 | 2.66 | 2.35 |
| Peak Time | 5.82 | 5.09 | 9.55 | 4.65 | 3.36 | 4.62 | 3.18 |
| Peak Marginal Penetration | 33.88 | 31.95 | 140.07 | 34.40 | 24.49 | 29.38 | 20.70 |

***Table 4b: Using ten years (T=10) of training data***

| **Item to be Predicted** | **Method** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Estimation Mean** | **Last Observation Projection** | **Classical Bass** | **Meta-Bass** | **Augmented Meta-Bass** | **Functional Regression** | **Augmented Functional Regression** |
| T+1 | 10.83 | 5.37 | 4.02 | 8.17 | 8.17 | 4.17 | 4.70 |
| T +2 | 11.19 | 6.59 | 6.38 | 8.69 | 8.73 | 5.48 | 5.69 |
| T +3 | 11.56 | 7.25 | 8.43 | 10.90 | 10.83 | 6.31 | 6.11 |
| T +4 | 11.58 | 8.44 | 11.02 | 9.40 | 9.26 | 8.08 | 7.81 |
| T +5 | 11.65 | 9.72 | 11.96 | 11.48 | 11.11 | 9.07 | 8.40 |
| Takeoff | 3.61 | 2.93 | NA | 2.95 | 2.86 | 2.69 | 2.51 |
| Peak Time | 4.69 | 3.95 | 7.54 | 3.50 | 2.94 | 3.66 | 2.92 |
| Peak Marginal Penetration | 26.66 | 23.65 | 42.71 | 25.99 | 22.32 | 23.18 | 18.18 |

Note: All results, except those for Takeoff and Peak Time, have been multiplied by $10^3$.
Using an alternative metric for error, the Mean Squared Error, yields similar results

*Table 5:*

***Superiority of Augmented Functional Regression Over Other Models***

***Table 5a: Fraction of curves for which Augmented Functional Regression outperforms other models using five years of training data***

| **Item to be Predicted** | **Method** | | | | | |
|---|---|---|---|---|---|---|
| | **Estimation Mean** | **Last Observation Projection** | **Classical Bass** | **Meta Bass** | **Augmented Meta-Bass** | **Functional Regression** |
| T+1 | 0.89 | 0.68 | 0.50 | 0.83 | 0.84 | 0.52 |
| T+2 | 0.79 | 0.70 | 0.50 | 0.77 | 0.76 | 0.58 |
| T+3 | 0.77 | 0.72 | 0.53 | 0.72 | 0.73 | 0.68 |
| T+4 | 0.77 | 0.77 | 0.61 | 0.74 | 0.70 | 0.74 |
| T+5 | 0.79 | 0.80 | 0.64 | 0.75 | 0.74 | 0.78 |
| Takeoff | 0.63 | 0.63 | NA | 0.58 | 0.53 | 0.60 |
| Peak Time | 0.76 | 0.69 | 0.76 | 0.66 | 0.53 | 0.67 |
| Peak Marginal Penetration | 0.70 | 0.69 | 0.66 | 0.72 | 0.59 | 0.69 |

***Table 5b: Fraction of curves for which Augmented Functional Regression outperforms other models using ten years of training data***

| **Item to be Predicted** | **Method** | | | | | |
|---|---|---|---|---|---|---|
| | **Estimation Mean** | **Last Observation Projection** | **Classical Bass** | **Meta-Bass** | **Augmented Meta-Bass** | **Functional Regression** |
| T+1 | 0.87 | 0.51 | 0.33 | 0.72 | 0.74 | 0.36 |
| T+2 | 0.81 | 0.62 | 0.50 | 0.72 | 0.75 | 0.50 |
| T+3 | 0.79 | 0.63 | 0.56 | 0.70 | 0.70 | 0.57 |
| T+4 | 0.75 | 0.61 | 0.64 | 0.62 | 0.62 | 0.59 |
| T+5 | 0.67 | 0.61 | 0.62 | 0.62 | 0.62 | 0.60 |
| Takeoff | 0.67 | 0.60 | NA | 0.58 | 0.57 | 0.52 |
| Peak Time | 0.69 | 0.61 | 0.77 | 0.53 | 0.49 | 0.56 |
| Peak Marginal Penetration | 0.72 | 0.67 | 0.63 | 0.66 | 0.60 | 0.68 |

*Table 6:*

***Proportions of Each Type of Product within Each Cluster***

| Product-Type | Clusters | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Cable TV | 16.7% | 16.7% | 10.5% | 7.7% | 8.9% | 4.2% |
| CD Player | 8.3% | 16.7% | 18.4% | 15.4% | 10.3% | 2.0% |
| DVD Player | 8.3% | 16.7% | 44.7% | 5.1% | 20.5% | 2.7% |
| Internet PC | 58.3% | 36.1% | 10.5% | 7.7% | 8.2% | 5.6% |
| Satellite TV | 0% | 2.8% | 10.5% | 16.7% | 15.1% | 5.3% |
| Videotape Recorder | 8.3% | 11.1% | 5.3% | 7.7% | 2.1% | 0% |
| Camera | 0% | 0% | 0% | 0% | 0.7% | 1.8% |
| Color TV | 0% | 0% | 0% | 2.6% | 0.7% | 2.4% |
| Fax | 0% | 0% | 0% | 3.8% | 2.1% | 0% |
| HiFi Stereo | 0% | 0% | 0% | 2.6% | 1.4% | 7.6% |
| PC | 0% | 0% | 0% | 2.6% | 4.8% | 8.9% |
| Telephone | 0% | 0% | 0% | 0% | 1.4% | 2.7% |
| Video camera | 0% | 0% | 0% | 2.6% | 4.8% | 1.8% |
| Videogame consol | 0% | 0% | 0% | 19.2% | 11.0% | 2.2% |
| Air-conditioner | 0% | 0% | 0% | 0% | 0.7% | 12.9% |
| Freezer | 0% | 0% | 0% | 0% | 0% | 7.1% |
| Microwave oven | 0% | 0% | 0% | 3.8% | 4.1% | 11.1% |
| Tumble Drier | 0% | 0% | 0% | 1.3% | 1.4% | 3.6% |
| Vacuum cleaner | 0% | 0% | 0% | 0% | 0% | 5.1% |
| Washing Machine | 0% | 0% | 0% | 1.3% | 0.7% | 2.0% |

*Table 7:*

***Distribution of Each Economic Grouping over Clusters***

| **Economic Groupings** | **Clusters** | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| N. America, W. Europe and Australasia | 5.1% | 10.8% | 13.6% | 15.9% | 23.3% | 31.2% |
| Eastern Europe | 0% | 3.4% | 3.4% | 8.0% | 24.7% | 60.3% |
| East Asia | 5.1% | 8.5% | 3.4% | 20.3% | 13.6% | 49.2% |
| West Asia | 0% | 6.7% | 11.1% | 13.3% | 22.2% | 46.7% |
| South America | 0% | 0% | 0.9% | 10% | 23.6% | 65.5% |
| Africa | 0% | 0% | 0% | 3.1% | 10.9% | 85.9% |
| Developing Asia | 0% | 2.3% | 0% | 3.8% | 8.3% | 85.6% |

***Figure 1:***

***Flowchart of the Implementation of the three FDA Techniques***

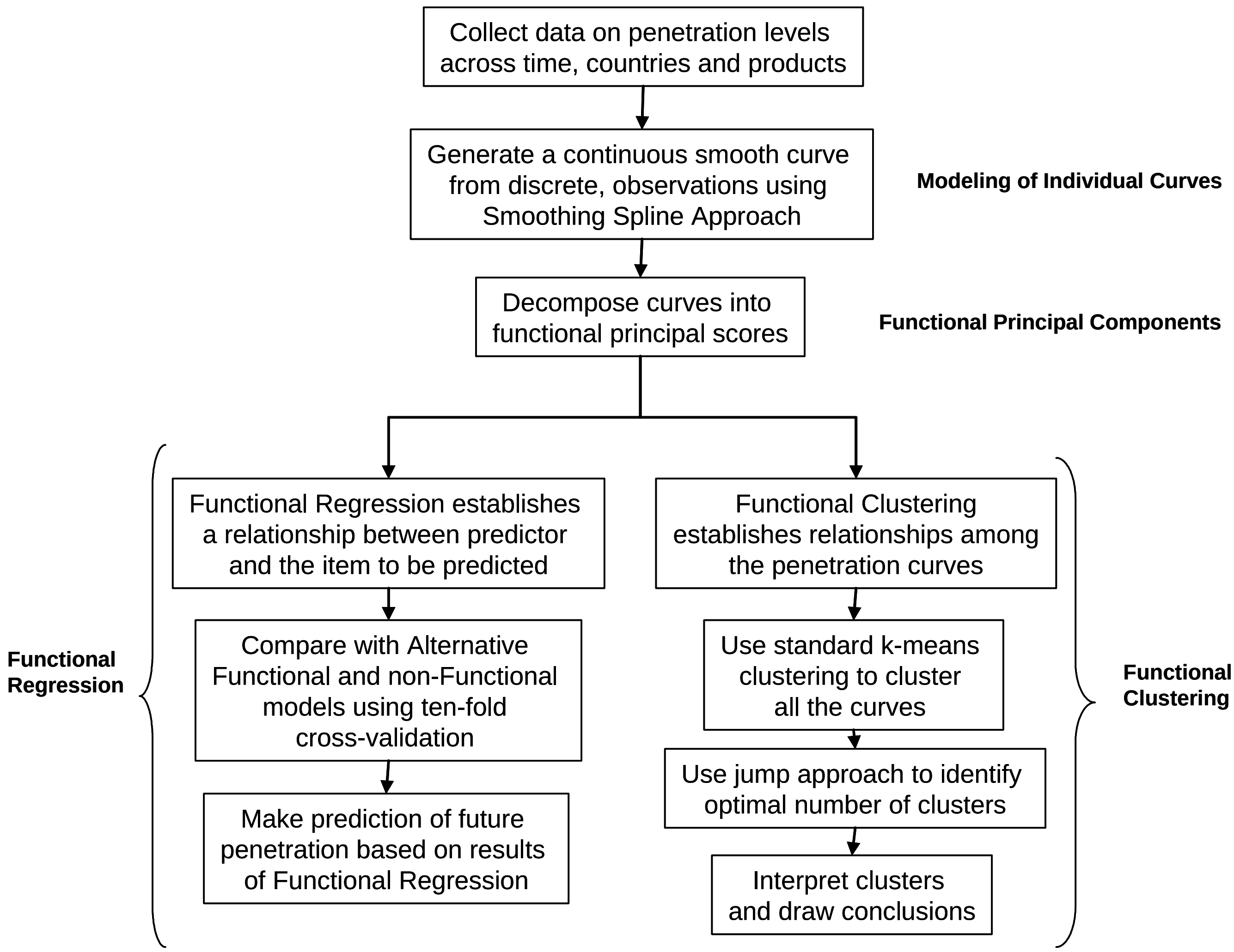


*Figure 2:*

***Analytic Framework demonstrating 10-fold cross validation***

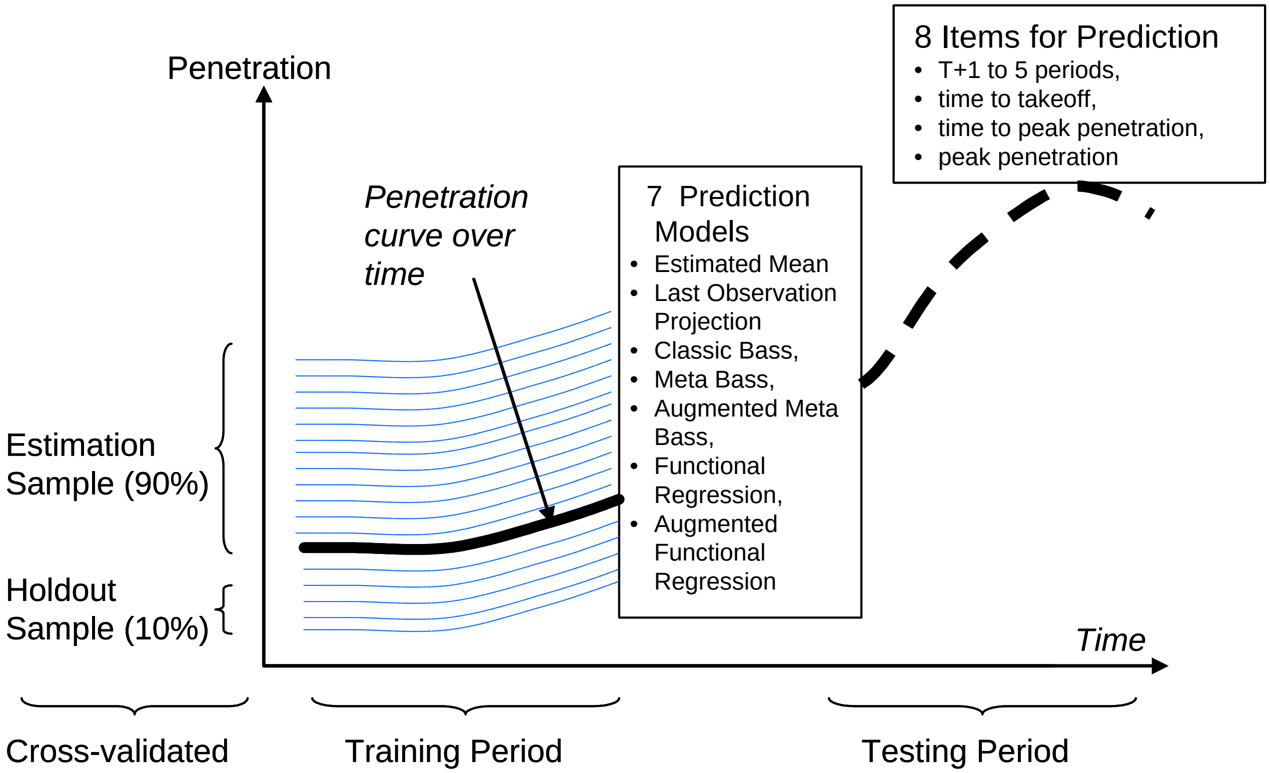
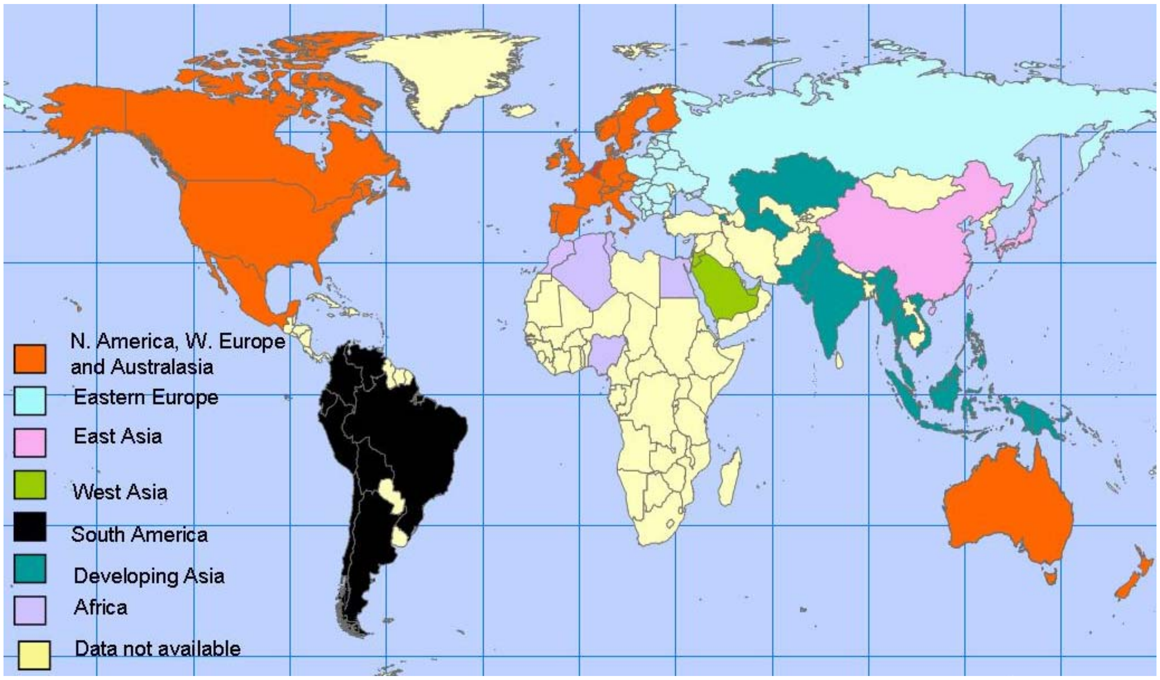


*Figure 3:*

*Distribution and Classification of Countries*

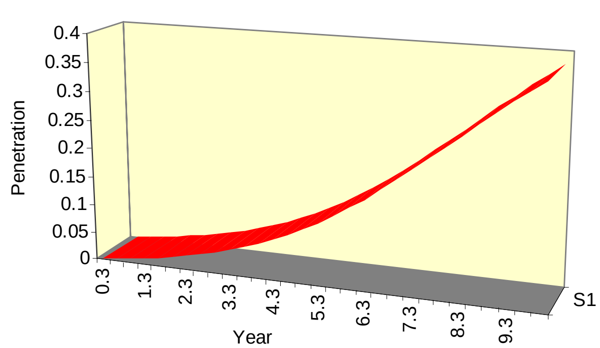



***Legend***

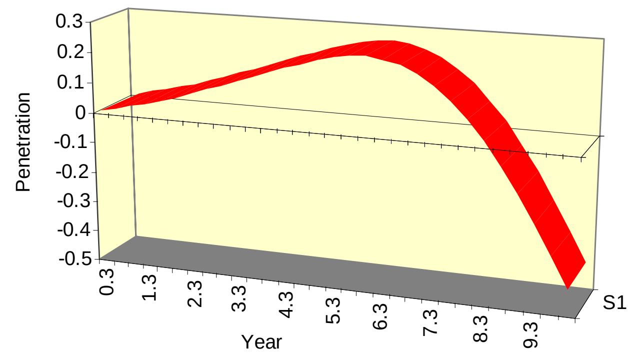
| Cluster | Countries |
|---|---|
| North, America, Western Europe, and Australasia | Canada, USA, Mexico, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, UK, Australia, N., Zealand |
| Eastern Europe | Belarus, Bulgaria, Croatia, Czech-Rep, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Russia, Slovakia, Slovenia, Ukraine |
| East Asia | China, Hong, Kong, Japan, S., Korea, Singapore, Taiwan |
| West Asia | Israel, Jordan, Kuwait, S., Arabia, UAE |
| South America | Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Peru, Venezuela |
| Africa | Algeria, Egypt, Morocco, Nigeria, Tunisia |
| Developing Asia | Azerbaijan, India, Indonesia, Kazakhstan, Malaysia, Pakistan, Philippines, Thailand, Turkmenistan, Vietnam |

## *Figure 4:*

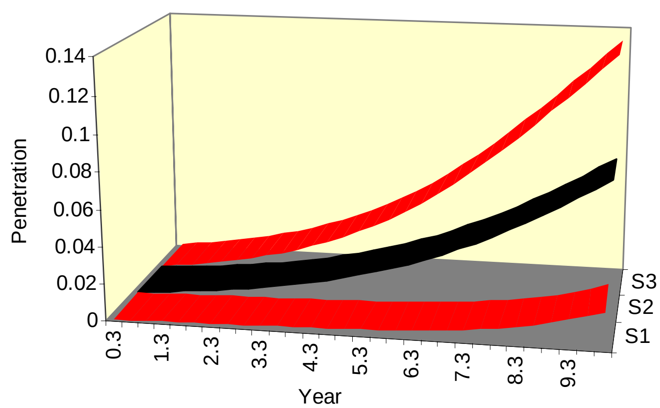
### *Illustration of First Two Functional Principal Component Curves (Based On Ten Years Of Training Data)*



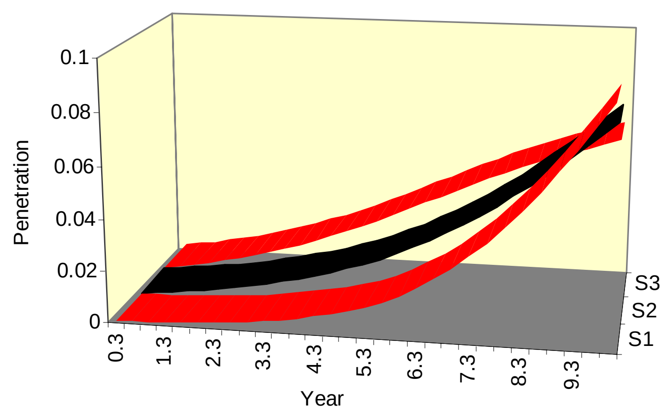

**PC1**




**PC2**




**Mean and PC1**




**Mean and PC2**

***Figure 5:***

***Illustration of Functional Clustering***

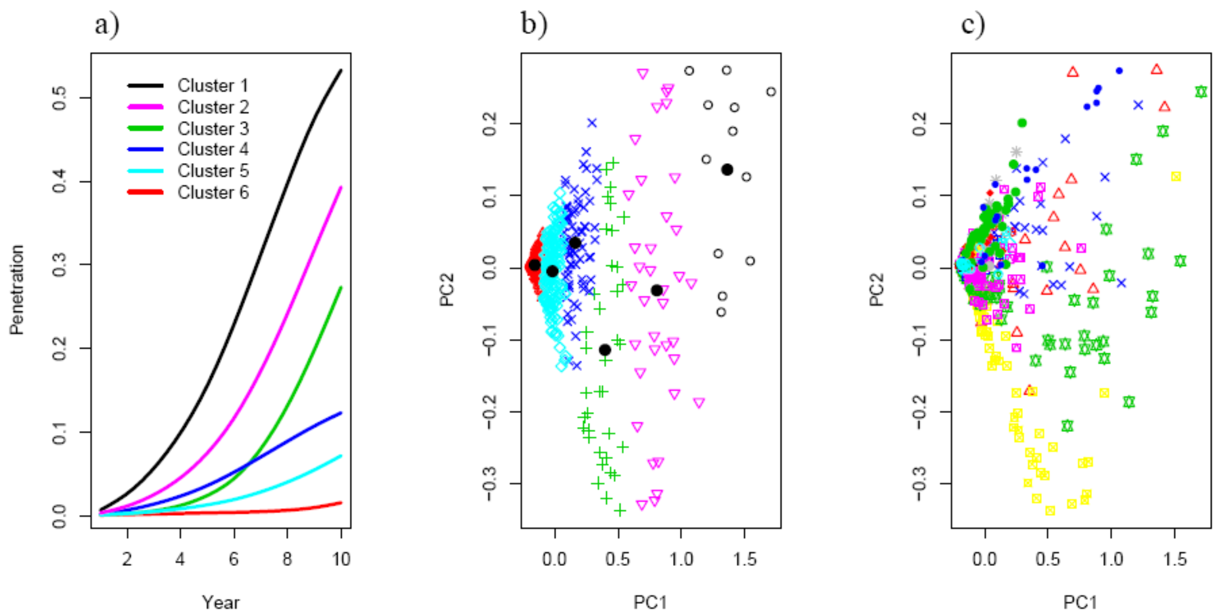
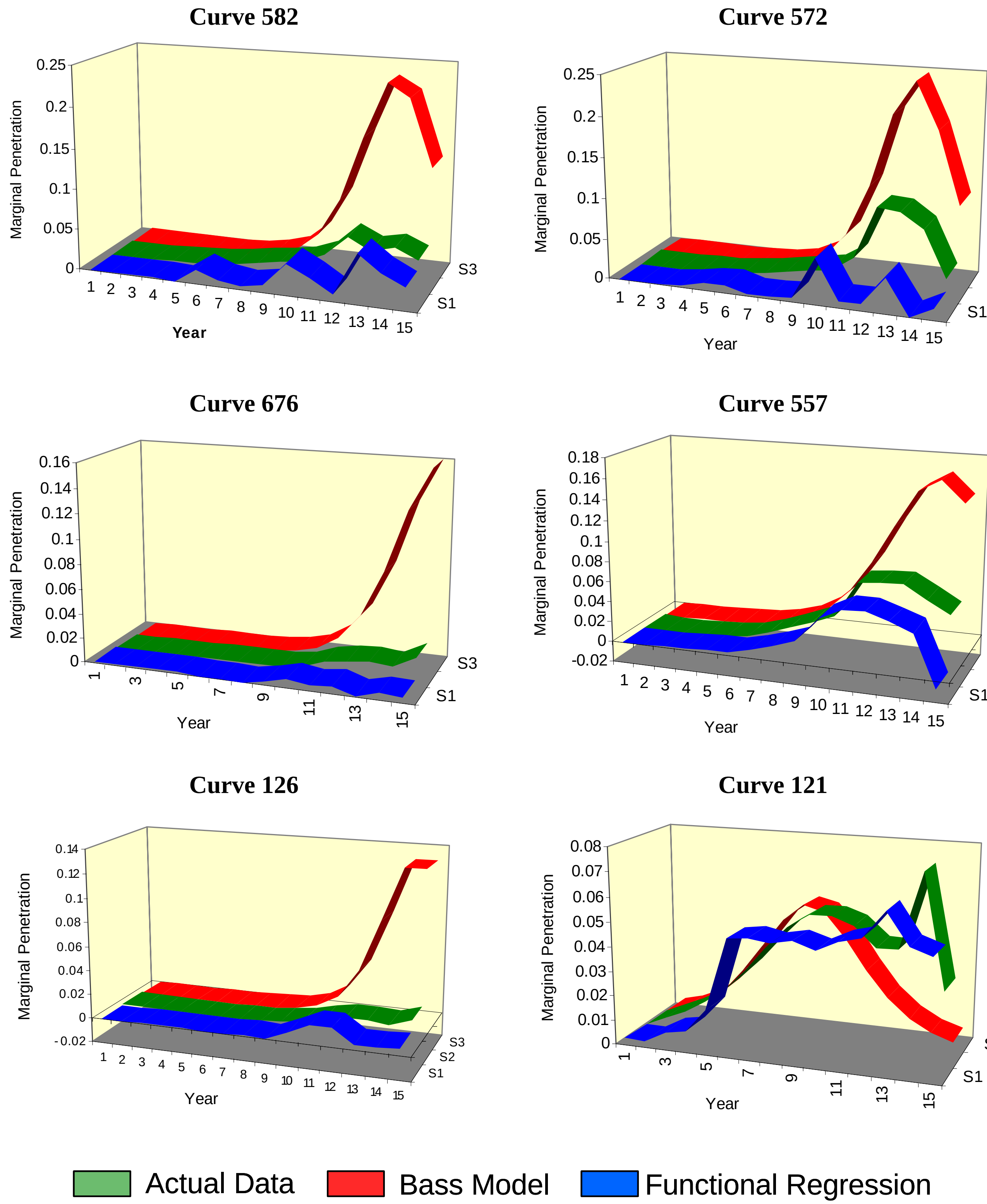



Figure 5 a) The shapes of the average penetration curves within each of the six clusters. b) The first two principal component scores for all 760 curves. A different color and plotting symbol has been used for each cluster with a black solid circle for the cluster centers. c) Same as b) but with different symbols for each product.

***Figure 6***

***Comparison of Predictive Accuracy of Classic Bass Model and Functional Regression Model***



# References

Agarwal, Rajshree and Barry L. Bayus (2002) "Market Evolution and Sales Takeoff of Product Innovations" *Management Science*, 48 (8), 1024–1041.

Armstrong, J. Scott (1978), "Forecasting with econometric methods: Folklore versus Fact," *Journal of Business*, 51, 4, 549-564.

____ (1984), "Forecasting by Extrapolation: Conclusions from 25 Years of Research," Interfaces, 14, Nov-Dec, 52-66.

____ and Edward J. Lusk (1983), "The accuracy of alternative extrapolation models: Analysis of forecasting competition through open peer review," *Journal of Forecasting*, 2, 3, 259-262.

Bass, F. M. (1969), "A new product growth for model consumer durables," *Management Science* 15, 215–227

Chandrasekaran, Deepa and Gerard J Tellis (2007), "Diffusion of New Products: A Critical Review of Models, Drivers, and Findings, *Review of Marketing,* 39-80.

____ and ____ (2008), "Global Takeoff of New Products: Culture, Economics or Vanishing Differences" forthcoming, *Marketing Science*.

Dekimpe, Marnik, Philip Parker, and Miklos Sarvary. (1998) "Staged Estimation of International Diffusion Models: An Application to Global Cellular Telephone Adoption." *Technological Forecasting and Social Change*, 57, 105–132.

Fader, P. S. and Hardie, B. G. 2005, "The value of simple models in new product forecasting and customer-base analysis: Research Articles," *Applied Stochastic Modeling Bus. Ind*. 21, 4-5 (Jul. 2005), 461-473.

Ganesh, Jaishankar, and V. Kumar. (1996) "Capturing the Cross-National Learning Effect: An Analysis of an Industrial Technology Diffusion." *Journal of the Academy of Marketing Science*, 24 (4), 328–337

____, ____, and V. Subramaniam, (1997) "Learning Effect in Multinational Diffusion of Consumer Durables: An Exploratory Investigation." *Journal of the Academy of Marketing Science*, 25 (3), 214–228.

Gatignon, Hubert, Jehoshua Eliashberg, and Thomas S. Robertson, (1989) "Modeling Multinational Diffusion Patterns: An Efficient Methodology." *Marketing Science*, 8 (3), 231–247.

Goldenberg, Jacob, Barak Libai, and Eitan Muller (2001) "Using Complex Systems Analysis to Advance Marketing Theory Development: Modeling Heterogeneity Effects on New Product Growth through Stochastic Cellular Automata." *Academy of Marketing Science Review*

Golder, Peter N., Gerard J. Tellis. (1997) "Will It Ever Fly? Modeling the Takeoff of Really New Consumer Durables" *Marketing Science* 16 (3), 256–270.

____ and ____ (1998) "Beyond Diffusion: An Affordability Model of the Growth of New Consumer Durables." *Journal of Forecasting*, 17 (3/4), 259–280.

____ and ____. (2004) "Going, Going, Gone: Cascades, Diffusion, and Turning Points of the Product Life Cycle." *Marketing Science*, 23 (2), 207–218.

____ (2000), "Historical Method in Marketing Research with New Evidence on Long-term Market Share Stability" *Journal of Marketing Research* 37 (May) 156-72.

Green, P. J. and Silverman, B. W. (1994), "Nonparametric Regression and Generalized Linear Models", Chapman and Hall

Hastie, T. J. and Tibshirani, R. J. (1990), "Generalized Additive Models", Chapman and Hall.

____., ____ and Friedman, J. (2001), "The Elements of Statistical Learning: Data Mining, Inference and Prediction" Springer.

Hauser, J. R., G. J. Tellis, A. Griffin. (2006), "Research on innovation: A review and Agenda for Marketing Science," *Marketing Science* 25(6) 687–717.

Helsen, Kristaan, Kamel Jedidi, and Wayne DeSarbo. (1993) "A New Approach to Country Segmentation Utilizing Multinational Diffusion Patterns" *Journal of Marketing*, 57 (4), 60–71

Jank W., and Shmueli, G., (2006) "Functional Data Analysis in Electronic Commerce Research", *Statistical Science*, vol. 21, no. 2, pp. 155-166.

James, G. M., Hastie, T., and Sugar, C. A. (2000), "Principal Component Models for Sparse Functional Data", *Biometrika* 87, 587-602

Kohli, Rajeev, Donald Lehmann, and Jae Pae, (1999) "Extent and Impact of Incubation Time in New Product Diffusion" *Journal of Product Innovation Management*, 16, 134–144.

Kumar, V., Jaishankar Ganesh, and Raj Echambadi (1998) "Cross National Diffusion Research: What Do We Know and How Certain Are We?" *Journal of Product Innovation Management*, 15, 255–268.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Mahajan, Vijay, Eitan Muller, and Frank M. Bass (1995) "Diffusion of New Products: Empirical Generalizations and Managerial Uses." *Marketing Science*, 14 (3), Part 2 of 2, G79–G88.

____, Eitan Muller, and Yoram Wind. (2000) “New Product Diffusion Models: From Theory to Practice.” In *New Product Diffusion Models*, ed. Mahajan, V., Eitan Muller, Yoram Wind, pp. 3-24, Boston: Kluwer Academic.

Peres, Renana, Eitan Mueller and Vijay Mahajan (2007), “Review of Research on Diffusion,” forthcoming, *Marketing Science*.

Putsis, William P. Jr., Sridhar Balasubramanian, Edward Kaplan, and Subrata Sen. (1997) “Mixing Behavior in Cross-Country Diffusion” *Marketing Science*, 16 (4), 354–369.

Ramsay, J.O. and Silverman, B.W. (2005), “Functional Data Analysis”, Springer (2nd Edition)

Rettinger, F., Jank, W., Tutz, G., and Shmueli, G., (2007) “Smoothing Sparse and Unevenly-Sampled Curves using Semiparametric Mixed Models: An Application to Online Auctions”, *Journal of the Royal Statistical Society*, Series C, accepted.

Stremersch, Stefan, and Gerard J. Tellis. (2004), “Managing International Growth of New Products” *International Journal of Research in Marketing.*

Sood, Ashish and Gerard J. Tellis (2005), “Technological Evolution and Radical Innovations,” *Journal of Marketing*, 69, 3 (July), 152-168.

Sugar, C. A. and James, G. M. (2003), “Finding the number of clusters in a data set: An information theoretic approach,” *Journal of the American Statistical Association* 98, 750–763.

Sultan, Fareena, John U. Farley, and Donald R. Lehmann (1990), “A Meta-Analysis of Diffusion Models*” Journal of Marketing Research*, 27, 70–77.

Talukdar, Debabrata, K. Sudhir, and Andrew Ainslie (2002) “Investigating New Product Diffusion Across Products and Countries” *Marketing Science*, 21 (1), 97–114.

Tellis, Gerard J and Philip Hans Franses (2006), “Optimal Data Interval for Advertising Response Models,” *Marketing Science*, 25, 3, May-June.

____, Stefan Stremersch, and Eden Yin (2003) “The International Takeoff of New Products: The Role of Economics, Culture and Country Innovativeness” *Marketing Science*, 22 (2), 188–208.

Van Den Bulte, Christophe, and Stefan Stremersch (2004) “Social Contagion and Income Heterogeneity In New Product Diffusion: A Meta-Analytic Test” *Marketing Science*, 23 (4), 530–544.

____, Yogesh V. Joshi. (2007), “New Product Diffusion with Influentials and Imitators”, *Marketing Science* 26(3) 400–421.

____. (2000) “New Product Diffusion Acceleration: Measurement and Analysis” *Marketing Science*, 19 (4), 366–380.

Venkatesan, R., Krishnan, T. V., and V., K. (2004), “Evolutionary estimation of macro-level diffusion models using genetic algorithms: An alternative to nonlinear least squares,” *Marketing Science* 23, 451–464. 12.