Universiteit Leiden

ICT in Business and the Public Sector

Applicability of diffusion models to forecast sales of Health Tech products

Name: David Vergara Manrique
Student-no: s1958046

Date: 12/10/2018

1st supervisor: Xishu Li
2nd supervisor: Thomas Bäck

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Innovation is a crucial concept for any company, being a proof of that the $500 billion that were invested in innovation in the US during 2015. For health tech companies, it is specially relevant, as this is a business industry which is constantly growing and where develop the last technology is needed to stay ahead of the competition.

The New Product Development (NPD) is the key process that enables companies to ideate new products and launch them into the market. In this particular process, the forecast of the number of future sales is a challenging but needed step, as it will allow the company to know how the volume of sales for a specific product will be spread over the time, and therefore, it will allow to plan and develop the needed strategies to keep and increase its market share. For this purpose, diffusion models seem to be the most popular tool used among health tech companies.

The aim of this particular research is to introduce a forecasting diffusion model into a leading health tech company, which is not currently using any diffusion model based on data. In order to do that, the Bass diffusion model is applied based on using historical sales data, concluding that is possible to apply this model to a health company with high accuracy. Finally, from this study it is also observed how is preferred to apply the Bass diffusion model on a one on one basis, generating the Bass model parameters out of a similar product to the one to be forecasted, rather than creating a general model based on several products, as the accuracy decreases.

Acknowledgements

First of all I would like to express my gratitude to my first thesis supervisor Xishu Li, for his continuous guidance, feedback and patient during the whole thesis process. I also want to thank my second supervisor Thomas Bäck, he provided me really good insights from a data science point of view, and he offered me feedback even during the busy times.

I would also like to thank my supervisors in the company, Chris and Alexei. They both showed a great interest on my thesis since the very beginning and provided me with data and all the resources I needed during the research. Their feedback was also really useful and allowed me to better understand the topic from a business perspective.

Last but not least, I would like to thank all my friends that have supported me during this process, and specially my parents, who have always been there providing me all the support and advice whenever I needed.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem statement

Innovation and the development of new products play a crucial role in all the producing companies, but especially in high tech companies, where the competition is really high and there exists a constant need for developing new products and improving the existing product portfolio. Innovation has been a recurrent topic in the literature, and it has been confirmed how what companies need the most for growing is innovation (Carden, 2005). Furthermore, it has been highlighted how important research and development (R&D) is in the industry, and the big investments those companies have to face in order to position themselves as market leaders (Coad & Rao, 2008). At the same time, it is emphasized how strong the competition is in the sector, and how difficult the process of developing a new product is, especially when it comes to forecasting future product sales. This forecasting difficulties arise, as in most of the cases, innovative products are unique in the market, and therefore it is hard to compare them to any other existing product.

The process of developing and introducing successfully a new product into the market is called New Product Development (NPD), for which 7 main activities are needed (Booz, Allen, & Hamilton, 1982): new product strategy, idea generation, screening and evaluation, business analysis, design and development, testing and commercialization; see Figure 1.

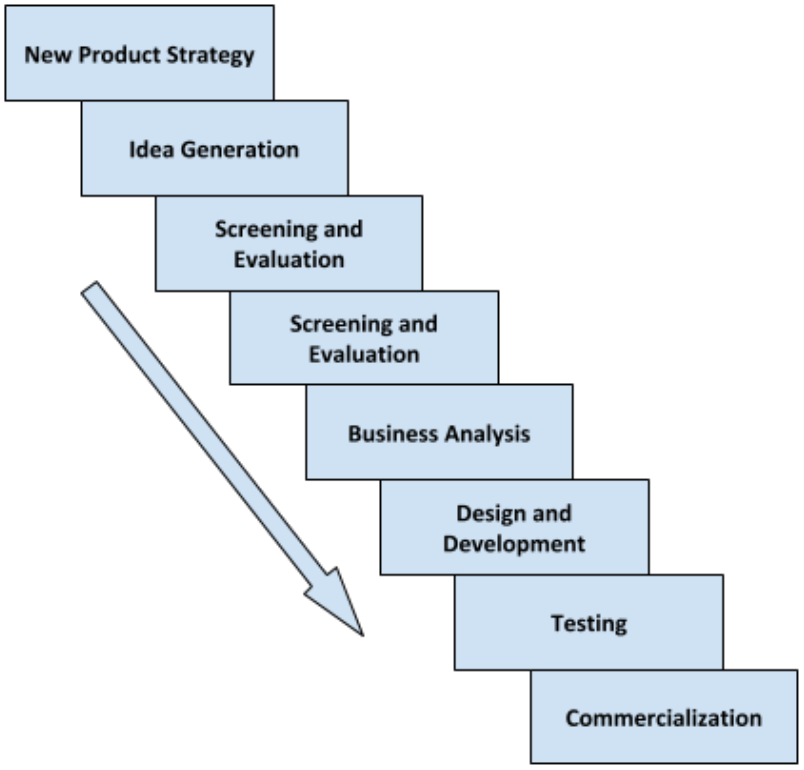


Figure 1: Booz, Allen, & Hamilton new product development model.

Among the 7 activities, business analysis is one of the most challenging ones, as it involves different quantitative market analyses, such as profit calculations, return on investment (ROI) estimation or forecasting of the sales volume. This thesis focus on this last type of analysis, for which different sales forecasting techniques are used in the high tech industry (Decker & Gnibba Yukawa, 2010), such as business cases analysis (the methodology currently used by the health tech company involved in this research), where different experts are consulted and different scenarios are built in order to predict sales; utility-based approaches, where the individual consumer behavior is studied (Erdem, Keane, & Strebel, 2005), or diffusion model approaches, which base the forecasts on the interpersonal communications among customers and on the cumulative past sales. Among diffusion models, the Bass forecasting diffusion model is the most relevant in the literature (F. Bass, 1969).

Diffusion models are the most common forecasting tool used in the high tech industry (Decker & Gnibba Yukawa, 2010), and several researches have focus on the adoption of this tool (Urban & Hauser, 1993) or (Wright, Upritchard, & Lewis, 1997). However, it is also possible to observe how there is a lack of application of diffusion models for the specific health tech industry, as most of the researches focus on generating different diffusion models estimates for traditional high tech products such as TVs (Jiang, Bass, & Bass, 2006). Furthermore, and apart from the great acceptance diffusion models present in the industry, some authors have questioned the validity of the diffusion models for sales forecasting in this kind of companies such as (Grantham, 1997) or (Jun, Kim, Park, Park, & Wilson, 2002).

Therefore, the aim of this research is to study the applicability of a diffusion model in a health tech company, with the main interest of the company being to replace the existing forecasting approach based on business cases by a data driven one. Furthermore, from a literature point of view, it will be interesting to understand how well a diffusion model performs in a health company and whether it is a good choice for sales forecasting.

1.2 Research questions

The main research question is:

How to use a diffusion model to forecast sales in a health tech company?

In order to address the main research question, three sub-questions will be answered:

- How accurate is to create a general set of parameters for the diffusion model in order to forecast sales of different products?
- How accurate is to generate an individual set of parameters for each product?
- How accurate is to use the same parameters for the diffusion model in different markets?

1.3 Thesis outline

This research is structured in 7 chapters as follows:

- Chapter 1 - Introduction: this chapter presents the problem statement, the different research questions to be answered in the thesis and the thesis outline.
- Chapter 2 – Literature Review: it defines the different diffusion model alternatives available in the literature and the specific diffusion model to be applied in the research.
- Chapter 3 - Methodology: it presents the methodology applied in this study in order to build the sales forecasting model.
- Chapter 4 – Analysis: it introduces the data provided by the company, the different preprocessing performed on it, and the algorithm used in order to build the forecasting model.
- Chapter 5 – Results: it shows the results of applying the model built in different scenarios and provides two different validations, against literature and against company estimates.
- Chapter 6 - Conclusions: it presents the conclusions drawn from this thesis, the research limitations that were faced during its development and future lines of research.

2 Literature Review

The diffusion model can be defined as one of the tools used in marketing in order to forecast the sales volume over the life-cycle of a new product (Mahajan, Muller, & Wind, 2000). Therefore, this literature review will first focus on the product life cycle concept, the key concept in relation with diffusion models, and then on the different diffusion models present in the literature. Finally, a diffusion model to be applied to the health tech company will be selected.

2.1 Product life cycle

Product life cycle is a concept traditionally associated to marketing, for which the first references in the literature can be found in the 1950s (Cao & Folan, 2012), whereas the main theories about this concept are established in the 1960s. Among the different theories and definitions proposed in those ages, the most relevant is the one presented by (Levitt, 1965), which is still used nowadays.

This theory explains how the sales curves presented by a product since it is introduced to the market, until it is discontinued, can be represented by a simple parabola. This parabola can be divided into four stages as shown in Figure 2. The first stage corresponds to the market development when the product is introduced into the market after proving that there is an existing demand for it; this stage presents low sales and a slow growth. The second stage is called growth, and it is characterized by a market expansion period which is followed by an increase in the number of sales. Following, the third stage called market maturity happens. This stage is defined by a deceleration in the fast growth and a stabilization in the number of sales, being typically the longest stage. It is at the end of this stage when new market opportunities are explored in order to introduce a new product. Finally, in the last stage called market decline, sales decrease rapidly ending in the discontinuation of the product. Furthermore, this parabola can be represented by the equation (1) (Cox, 1967), which is used as the main equation in several diffusion models.

$$Y = a + bX + cX^2 \tag{1}$$

Where:

Y : sales

X : time

a, b, c : aggregation parameters

Recent studies, focus on how product life cycle models can be used to compare different companies and different market strategies by studying their behaviours over time (Werker, 2003). This can be used in order to create insights for the management decision making processes, such as the one happening during the new product development. On the other hand, other studies emphasize on how broad the research activities in the product life cycle area are, and point towards those areas

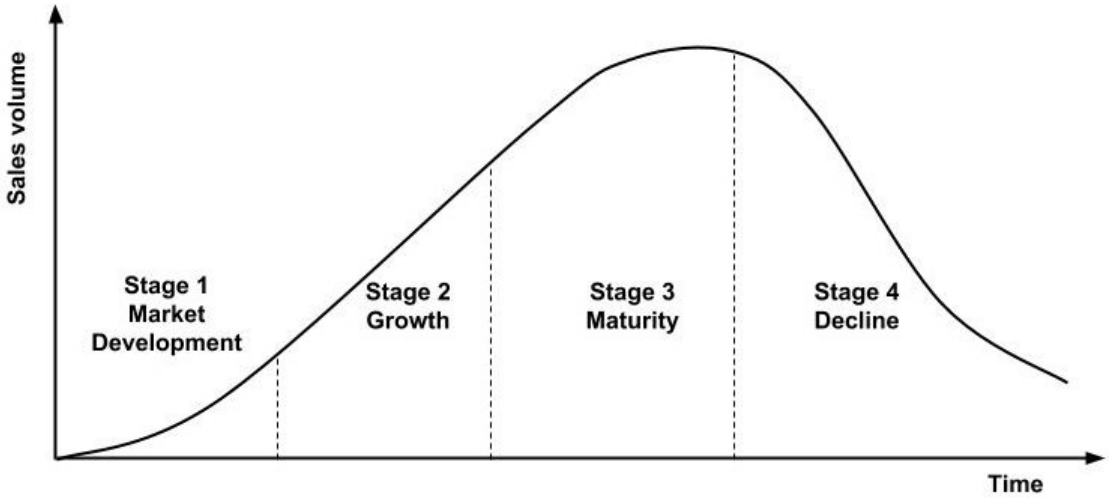


Figure 2: Typical product life cycle curve as described in the literature (Levitt, 1965).

where further development is needed (Rink & Swan, 1979). One of those areas, especially relevant for this research, is the forecasting of product life cycle stages. Most of the methods focus on forecasting the sales numbers within the next product life cycle stage, based on those of previous stages, which implies a non-well anticipated and not accurate forecasting. Therefore, this research contributes to this area by forecasting the whole product life cycle based on historic sales of similar products, before the first stage has started.

2.2 Diffusion models

The diffusion of innovations concept can be defined as the process of communicating an innovation to the members of a social system over time, through one or more communication channels. (Rogers, 1995). One of the applications of the previous definition was stated by (Mahajan & Muller, 1979), where they described diffusion models as those models that aim to show how an innovation spreads among a group of adopters over time, focusing on the development of a product life cycle parabola that serves as an indicator of when the first purchases of the innovation are happening.

Several authors split the different diffusion models in two main categories: basic diffusion models and extended diffusion models (Jaakkola, 1996) or (Mahajan, Muller, & Bass, 1990). On the one hand, basic diffusion models are the classical diffusion models that were mainly introduced between the 1960s and 1980s. Those diffusion models are characterized by using a small number of parameters and by not including decision variables. Furthermore, basic diffusion models can be divided into those using aggregate market data, looking at the whole market behaviour, and the ones using individual market data, which base the forecasting on decisions made by individuals within the market. On the other hand, the extended diffusion models are those generally introduced after 1980. They are characterized by using a basic diffusion model as reference, and extending it by including decision making variables or marketing mix variables; and, therefore, increasing

its complexity. In the following sections, different diffusion models for the different categories are introduced.

2.2.1 Basic aggregate diffusion models

One of the most popular basic diffusion models is the **Fisher and Pry model** (Fisher & Pry, 1971). It is based on three assumptions: 1. Old technology will be replaced by new one, which therefore can be considered as a competitive substitution. 2. If a substitution progresses beyond a certain threshold, it will penetrate the market until completion. 3. The fractional rate of new adopters is proportional to the remaining amount of old technology users to be substituted. Those three assumptions are based on the imitation effect, which states that individuals will buy a product if other individuals have already bought the product, as they will be influenced by the word of mouth or internal communication effect. Therefore, if a product has reached a certain threshold, and according to the model, the world of mouth will be strong enough to penetrate the product into the market. It is defined by the following equation (2):

$$f = \frac{1}{1 + e^{-b(t-t_0)}} \tag{2}$$

Where:

f : market percentage that have adopted the new product

b : potential growth (imitation effect)

t : time since the product was introduced

Another popular basic diffusion model is the one developed by **Fourt and Woodlock** (Fourt & Woodlock, 1960). In contrast to the previous one, this model is based on the innovation effect, and therefore, uses it as the unique parameter. This effect states that people will buy a product as they are only influenced by external communication such as advertisement or mass media. Therefore, this method suggests that a strong market penetration effect will happen at the beginning of the product life cycle, and, later, the increments in penetration will be proportional to the remaining distance to the penetration ceiling, where the whole potential market has been fulfilled. The model equation is the following (3):

$$f_t = rM(1-r)^{t-1} \tag{3}$$

Where:

f : change in cumulative sales at time t

r : rate of penetration (innovation effect)

M : total potential buyers

The **Mansfield method** (Mansfield, 1961) is also based in the previously introduced imitation effect. In particular, Mansfield stated that when the number of firms adopting an innovation increases, less investments are required for develop further innovations and for the adoption of the existing innovations. This model generates the product life cycle curve by using the coefficient of imitation and the cumulative number of adopters for the current period of time as the unique parameters. It is defined by the following equation (4):

$$m_{ij}(t) = n_{ij}[1 + e^{C_{ij}t}]^{-1} \tag{4}$$

Where:

$m_{ij}(t)$: cumulative number of firms that have introduced the innovation at time t

n_{ij} : total number of firms

C : coefficient of imitation

Finally, the **Bass model** (F. Bass, 1969) was introduced as a combination of the two previous methods, including in its definition the innovation and imitation effects. It considers that potential adopters in a population are divided into two groups, the first one affected by mass media and the second one affected by word of mouth communication. Figure 3 shows how both diffusion patterns (innovation and imitation) affect the Bass model. It is possible to observe how innovators present a stronger weight at the beginning of the life cycle, and how imitators are leading during the rest of the life cycle. By combining both patterns, the general Bass diffusion model pattern is generated, corresponding to all the new adopters of the innovation.

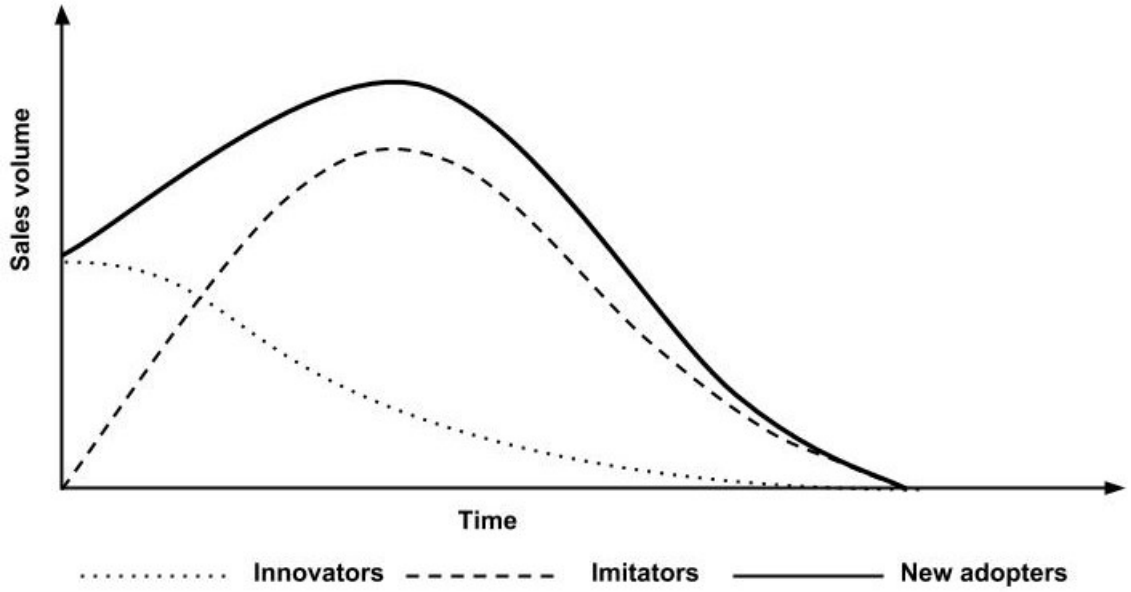


Figure 3: Diffusion patterns present in the Bass model as described in the literature (Massiani & Gohs, 2015).

Therefore, this model presents one parameter for the innovation effect p, one parameter for the imitation effect q and one parameter for the market potential m. It is considered one of the most used and studied methods, and it has been selected as one of the most frequently cited papers in the academic journal Management Science (Science, 2004). It can be expressed based on the following equation (5):

$$S(T) = pm + (q - p)Y(T) - (q/m)[Y(T)]^2 \tag{5}$$

Where:

$S(T)$: sales at time T

$Y(T)$: number of cumulative sales at time T

p : coefficient of innovation

q : coefficient of imitation

m : estimated number of sales during the whole product's life cycle

2.2.2 Basic individual diffusion models

The first individual diffusion model to be studied is the **Chatterjee and Eliashberg model** (Chatterjee & Eliashberg, 1989). On the one hand it focuses on individual level determinants of adoption: perception of the performance of the innovation, risk aversion, price sensitivity and responsiveness to new information about the innovation. All of them are obtained from the potential customers through a survey. On the other hand, true performance of the innovation and the real price are captured. Using the previous data, three parameters a,b and u are generated; a being an indication of how far is the consumer from adoption prior launch, b indicating the price hurdle for the consumer and u the real product performance. This way, a value of a below 0 would mean the adoption of the innovation when it becomes available; otherwise, the consumer will adopt as long as his price hurdle, value b, is smaller than the real performance, value u. It is defined by the following equation (6):

$$y = \frac{a}{u - b} \tag{6}$$

Where:

y : indication of product adoption, the smaller the earlier

a : indication of how far is the consumer from adoption prior launch

b : indication of the price hurdle for the consumer

u : the real product performance

A different individual diffusion model is the one presented by **Oren and Schwartz** (Oren & Schwartz, 1988). This model is mainly based on the risk aversion factor. It states that the customers that experience a small risk aversion are the ones adopting the innovation first. It uses five parameters to build the forecasting: the distribution on risk aversion in the population, the flow rate of consumers, the success rate for the current technology, the initial success rate for the new technology and uncertain success rate for the new technology. Those parameters are obtained by performing a market analysis. Furthermore, an interesting aspect of this model, is that if risk aversion for the adopters follows a negative exponential distribution, this model is reduced to the **Mansfield** aggregate model previously introduced. The model equations are the following (7)(8):

$$N(t) = a[1 - exp(-yk(N + N_0 + 1))] \tag{7}$$

$$k = \frac{2(\Theta_0 - \Theta_c)}{1 - \Theta_0} \tag{8}$$

Where:

y : distribution on risk aversion in the population

a : flow rate of consumers

Θ_c : success rate for the current technology

Θ_0 : initial success rate for the new technology

N : uncertain success rate for the new technology

Finally, the individual diffusion model introduced by **Lattin and Roberts** (Lattin & Roberts, 1989) is presented. This model is also based in the risk aversion concept, but in this situation, it includes a utility threshold to measure customer expectancy against innovation offerings. It uses five variables to provide the forecasting: time, the upper bound of the uniform distribution, the degree of risk aversion, the consumer utility for the new product and the consumer preference for the innovation under certainty. The authors suggested how this model can show a similar or even better performance than the **Bass** model, using for that two extra parameters. The model can be observed in the following equation (9):

$$N(t) = a + bN(t-1) - \frac{d}{c + N(t-1)} \tag{9}$$

Where:

N : adopters at time t

t : time

a : upper bound of the uniform distribution

b : degree of risk aversion

c : the consumer utility for the new product

d : consumer preference for the innovation under certainty

2.2.3 Extended diffusion models

The main quality of the extended diffusion models is the increase of accuracy and flexibility they provide to an existing basic model, by adding extra parameters, or how they combine several basic models to create a more robust and accurate model.

The first extended diffusion model to be presented was introduced by **Shafir and Kabir** (Sharif & Kabir, 1976). This model is based on 3 basic models, the previously introduced Fisher and Pry model (Fisher & Pry, 1971), the **Blackman model** (Blackman, 1972) and the **Floyd's model** (Floyd, 1962). This kind of models that are based on more than one basic model are also known as umbrella models. According to the authors, the reason for combining those 3 models is that the first two models, in general, produce too optimistic forecasts, whereas the third one produces too pessimistic ones. This way, the combination of them produces a more reliable and accurate model. As three basic models are combined, the complexity of the new model increases, involving in this situation 8 parameters: time, market share at time t, constant 1, constant 2, data scatteredness, data extent, last value of market share and effective life span. It is defined by the following equations (10)(11):

$$(1-\Theta)[ln\frac{f}{F-f}]+\Theta[ln\frac{f}{F-f}+\frac{F}{F-f}]=C_1+C_2t \tag{10}$$

$$\Theta=\phi[DS,DE,fl,ELS] \tag{11}$$

Where:

t : time

f : market share at time t

C_1 : constant 1

C_2 : constant 2

DS : data scatteredness

DE : data extent

fl : last value of market share

ELS : effective life span

Apart from combining different basic models, it is also possible to extend one existing basic model as it was explained before. One of the most extended basic models is the Bass model, and one of those extensions is the one performed by **Guo** (Guo, 2014). The Bass model, as most of the basic models, only allows to predict the first purchases made by a specific customer. In this approach, the Bass model is extended in order to also forecast the repeat purchases. For that purpose, the Novelty Loyalty Based Consumer Utility theory (Faison, 1977), which explains how novelty seeking and loyalty seeking instincts pervade the human behaviour, is included into the model. This way, two new parameters are added to the three already existing parameters in the Bass model: one scaling parameter and the novelty decay parameter. Those parameters are determined by the human biology, and therefore bio-psychological experiments will have to be conducted in order to calculate them, thereby increasing significantly the complexity of the model in terms of the applicability. The model can be observed in the following equations (12)(13)(14)(15):

$$U(T) = S(T) + RE(T) \tag{12}$$

$$S(T) = pm + (q - p)Y(T) - (q/m)[Y(T)]^2 \tag{13}$$

$$RE(T) = \int_0^T S(t)r(T,t)dt \tag{14}$$

$$r(T,t) = B(1 - a(T - t))(T - t) \tag{15}$$

Where:

$S(T)$: sales at time T

$Y(T)$: number of cumulative sales at time T

p : coefficient of innovation

q : coefficient of imitation

m : estimated number of sales during the whole product's life cycle

a : novelty decay parameter

B : scaling parameter

2.3 Models comparison

The different diffusion forecasting models that were discussed in the previous chapters are summarized in the Table 1. It is possible to observe how the aggregate models base their forecasting mainly on the innovation and imitation effect, how the individual models base the forecast on the risk aversion, and how the extended models present a combination of the previous ones. Furthermore, it is possible to see how the basic models are not as complex as the extended models in terms of parameters.

Model	**Description**	**Type**	**Parameters**
Fisher and Pry	Focus on the Imitation effect	Basic, aggregate	2
Fourt and Woodlock	Focus on the Innovation effect	Basic, aggregate	2
Mansfield	Focus on the Imitation effect	Basic, aggregate	2
Bass	Focus on a combination of imitation and innovation	Basic, aggregate	3
Chatterjee and Eliashberg	Focus on individual level determinants of adoption	Basic, individual	3
Oren and Schwartz	Focus on risk aversion. Mansfield model when negative risk aversion	Basic, individual	5
Lattin and Roberts	Focus on risk aversion	Basic, individual	4
Shafir Kabir	Combines 3 basic models, balance of optimistic and pessimistic models	Extended, Umbrella	8
Guo	Extended Bass model including repurchasing	Extended	5

Table 1: Overview of the diffusion models studied in the literature review

3 Methodology

3.1 Research strategy

The research performed during this thesis corresponds to an applied one, in the field of marketing and sales forecasting diffusion models. Several applied researches in regard to diffusion models have been performed in the literature before, however, most of them focus on high tech companies and traditional products such as TVs. The aim of this research is exploring how to apply a diffusion model to a leading health tech company, a topic that has not yet been explored in depth in the literature.

3.2 Research method

The research performed in this thesis is based on a deductive research approach in conjunction with a quantitative analysis in order to investigate the research questions stated in the Chapter 1. Furthermore, a literature review has been conducted in order to define the problem statement, and to study the particular diffusion model to be applied. The search engine used during the thesis development to carry out the literature review is Google Scholar (Google Inc, 2018), which has been queried by using the following key words: high tech, health tech, new product development, new product introduction, diffusion model, sales forecasting model and Bass model.

3.3 Data collection

On the one hand, historical sales data from the company has been acquired, in order to build a sales forecasting diffusion model based on this data. This data corresponds to sales of electric toothbrushes, the product that is going to be used to test the model, over the last 8 years and from two different countries: United States and China. In total, a data set with 15 variables and 30.000 entries was provided, from which 30 different products were extracted.

On the other hand, a structured survey has been used in order to obtain estimates from the product managers in the company. Those estimates focus on the product life cycle for the different products involved in this research. A structured survey has been selected as it is a tool that provides certain benefits such as low cost, high accuracy or small response time, when several answers need to be acquired.

3.4 Research process

The first meetings were held in May of 2018 with the company managers in order to get acceptance of the research and the use of company data. The whole research has been carried between May and October of 2018 and it took place in the headquarters of the company, in collaboration with the

data science department. Different meetings were held during this period, discussing the different analysis applied in the research with the company managers, at the same time a close supervision was provided by the University. Finally, the structured survey developed in order to acquire product life cycle estimates from the company was distributed in September of 2018, getting a good acceptance among the managers.

3.5 Data analysis

In order to structure the data analysis, the cross-industry standard process for data mining (Shearer, 2000) is followed. This standard defined by data mining experts, states an iterative process in which 6 stages are defined: business understanding (performed during the introduction and literature review, Chapters 1 and 2), data understanding (Chapter 3), data preparation (Chapter 4), modeling (Chapter 4), evaluation (Chapter 5) and deployment (to be implemented by the company managers). Finally, the programming language and software environment for statistical computing R (R Core Team, 2013) was used during the whole data and statistical analysis performed in this research.

3.6 Diffusion model

A final step that was needed to be performed before starting the analysis, is the selection of the diffusion model that is going to be applied in this research. As stated above, the purpose of this research is to apply a diffusion model to a health tech company which is currently using a business case approach, not based on data, for forecasting sales. Therefore, the main objective is to introduce a diffusion model able to produce good forecasts, but at the same time, easy to use in order to make its adoption simple. Therefore, among the aggregate level diffusion models introduce in the literature review chapter, it is easy to see the Bass model as the best option, as it combines the imitation and innovation effects, and it provides a better accuracy.

When comparing the chosen aggregate model (Bass) with the individual models, the decision of choosing a model is not easy to make, as individual models are also easy to use and some of them suggest having a better performance than the Bass model, such as (Lattin & Roberts, 1989). However, and according to (Mahajan, Muller, & Bass, 1993), individual models do not present better results than aggregate models in general, and they present problems in long term forecasting, something relevant for a health teach company. Therefore, the Bass model seems to be a better option than the individual models.

Finally, the first advantage when comparing the Bass model with the extended models is its simplicity. Furthermore, as stated by (F. M. Bass, Trichy, & Dipak, 1994), the Bass model is able to get as good estimates as extended models, without using any decision variable. Apart from that, the Bass model is the most influential diffusion model in the literature, presenting several extensions, which could be introduced by the company in a future stage of development. Therefore,

the Bass model seems to be the best option in terms of simplicity, accuracy and literature research.

4 Analysis

As stated above, the particular product used to test the Bass diffusion model in this research is the electrical toothbrush. The process of data preparation and analysis in this research is described in the following:

4.1 Data preparation

Sales of toothbrushes over the last 8 years and from two different countries, United States and China, were extracted from the system. The following preprocessing is applied to the data:

Selecting data

First of all, different filters where applied in the product types, deleting the following entries: trial or promotion products that does not count as a sale; stickers and different types of merchandising; and toothbrush accessories, such as toothbrush cases or chargers. The reason for deleting those products is that the forecasting model that is going to be build aims for forecasting toothbrushes sales, without taking into consideration other types of products or product variations or accessories that would include noise into the model.

Then, out of the starting 15 starting variables, 11 variables were deleted for not being relevant for the research, as most of them represent internal company codes or product descriptions. The 4 remaining variables correspond to: product number, month, country and number of sales per month.

Grouping product data

Over 500 product variations are present in the data corresponding to different colors or different packaging where some accessories are included. In order to group that product data at the right level, a product definition needs to be provided. After consulting with a manager in New Product Introduction (NPI) which product level is considered during that process, the following product definition was generated:

"A product in the New Product Introduction (NPI) can be defined as an item that provides a specific and unique functionality regardless of the color, accessories, marketing or packaging"

Therefore, all the different color variations or different packages generated for selling the product (including traveling case, two charges, etc.) can be grouped based on functionality. After performing this grouping, 30 final products are taken into consideration, 16 for the United States market and 14 for the China market.

Data normalization

The method to be developed will not work properly if there is data that is not normalized, as further algorithms to be applied such as clustering will be affected by this feature. Therefore, in order to avoid this problem, two main variables need to be normalized, time and sales.

On the one hand, historic data from the last 8 years was retrieved, this is translated into having date entries from the first of January 2011 until the first of June 2018. For the forecasting model, only data grouped as a month level is needed. In order to normalize this measure, a new interval variable is introduced coding the months from 1 to 89, representing the number 1 the month January 2011 and the number 89 the month May 2018.

On the other hand, sales data is different from one product to another, being the average number of sales per month for some products in the range of thousands of units, whereas other products only sell hundreds of units. This is caused basically because different products focus on different markets segments. To solve this, the number of sales per product per month is divided by the total number of sales for that product, re-scaling this way the numerical data from 0 to 1.

Missing values

Missing values is another relevant topic to take into consideration as it could have a big effect on the model. There are two considerations regarding missing values that were detected in the data.

First, not all the products present sales for all the 89 months, mainly because some products were introduced later than 2011 or discontinued before 2018. In order to solve this problem, the sales for those specific months for the products involved are coded as "NA". This way it is possible to process data sets of 89 observations for all the products.

A second consideration regarding those products introduced after 2011 and not discontinued yet, is that their product life cycle is not complete yet. For this reason, those products will only be used for testing purposes and not for training the model.

Censoring data

The next step into the data cleaning is to apply censoring to the data, removing those data points that are only partially known and that insert noise into the model. In order to do this, box plots were used to detect this kind of data. It is possible to observe how in different products some data appears in form of a "long tail" of small sales at the end of the product life cycle. It can be observed in Figure 4.

After discussing with the managers, the reasons behind that "long tail" were clarified. This effect happens because the product life cycle of the products is artificially extended in the market, in order to sell the last available units or for internal policies reasons. As the interest of this research is to develop a forecasting model of the natural product life cycle of health tech products, this artificially generated tail is removed.

Derived new variables

The last step in this data preparation process is to calculate new variables in the data set that are needed to build the forecasting model. In total, six new variables are introduced, that will be used later in order to generate the forecasting model. Those variables are the already discussed normalized sales, normalized months (1-89), the cumulative number of normal sales and normalized sales, and the squared cumulative number of normal sales and normalized sales.

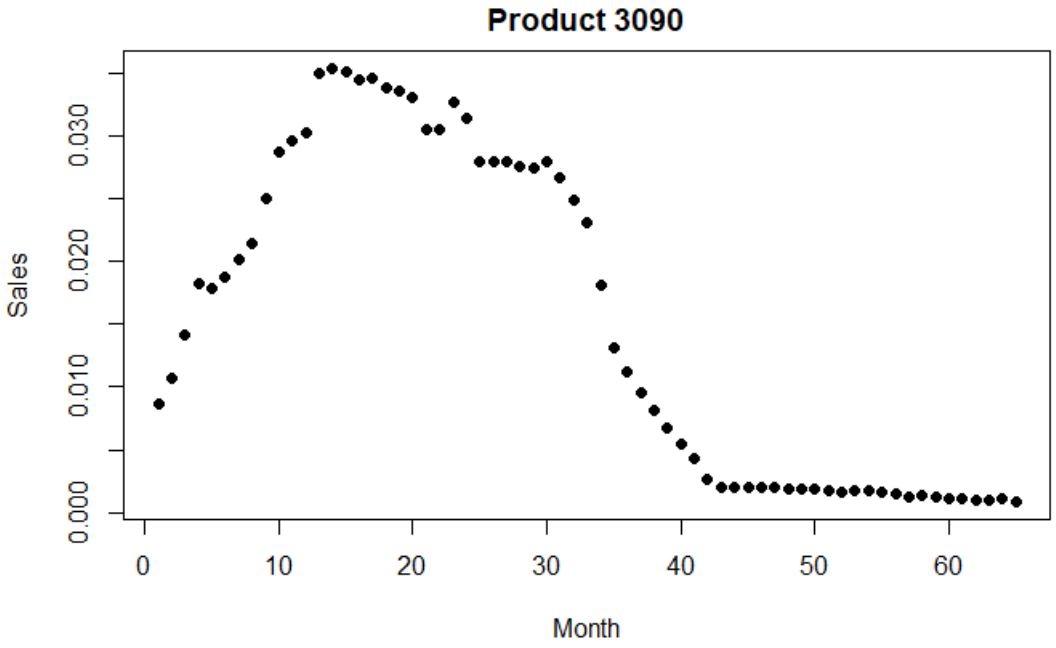


Figure 4: Sales per month for the product 3090.

4.2 Descriptive analytics

From the original data set involving 15 variables and 30.000 entries, a new data set has been generated involving 9 variables and 2315 observations. All those variables are described in Appendix A.

Sales for two countries are stored in the data set, United States and China, being 16 products present for the first country and 14 for the second one. Out of those products, 13 products are present in both countries and only 10 products out of the 30 have finished the product life cycle (products that do not present sales in the last 6 months). This is summarized in Table 2 and Table 3.

Country	Products	Finished Life Cycle	Not finished Life Cycle
US	16	7	9
China	14	3	11

Table 2: Products present in the different countries

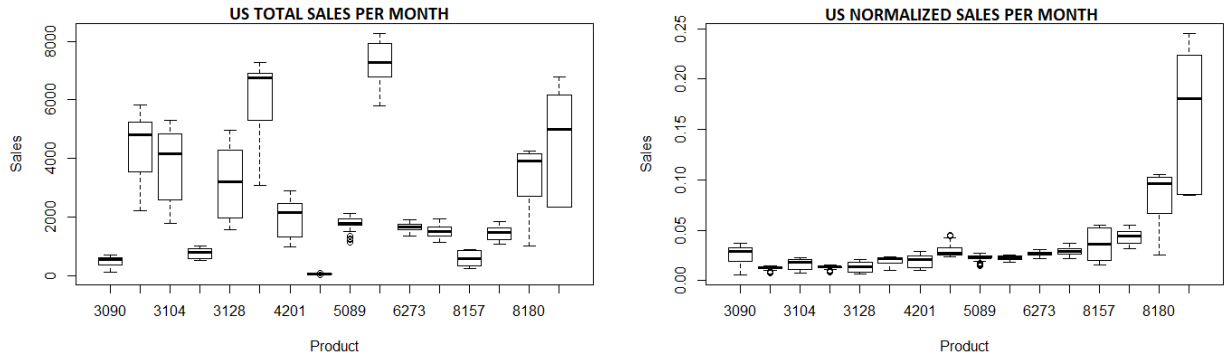
Furthermore, 4 box plots are generated in order to get a better understanding of the different sales per product in the US and China markets. Figure 5a shows a comparison of the sales for the different products in the US. It is possible to observe how the sales vary a lot between different products, for example products like 6272 selling in terms of 7000 units per month whereas other products like 3090 sells less than 1000 per month. This is caused because different products focus on different market segments, so these differences in number of sales among products are expected.

Product	US	CH	Finished Life Cycle US	Finished Life Cycle CH
3080	No	Yes	X	Yes
3090	Yes	No	Yes	X
3101	Yes	Yes	No	No
3104	Yes	Yes	Yes	No
3105	Yes	Yes	No	No
3128	Yes	Yes	Yes	No
3159	Yes	Yes	Yes	Yes
4201	Yes	Yes	Yes	No
4395	Yes	Yes	No	Yes
5089	Yes	Yes	Yes	No
6272	Yes	Yes	No	No
6273	Yes	Yes	Yes	No
7171	Yes	Yes	No	No
8157	Yes	Yes	No	No
8179	Yes	No	No	X
8180	Yes	Yes	No	No
8614	Yes	No	No	X

Table 3: Detailed products present in the different countries

Apart from that it is possible to see how most of the outliers were removed during the cleaning phase (only product 5089 presents some outliers that are close to the minimum).

On the other hand, looking at Figure 5b , the comparison of normalized sales for different products in the US, it is possible to see how after normalization most of the products present the same distribution. Only the last 2 products 8180 and 8614 look slightly different; this is caused because those are products recently introduced into the market, and therefore there is still a small number of data points for them.



(a) Sales for different products in the US

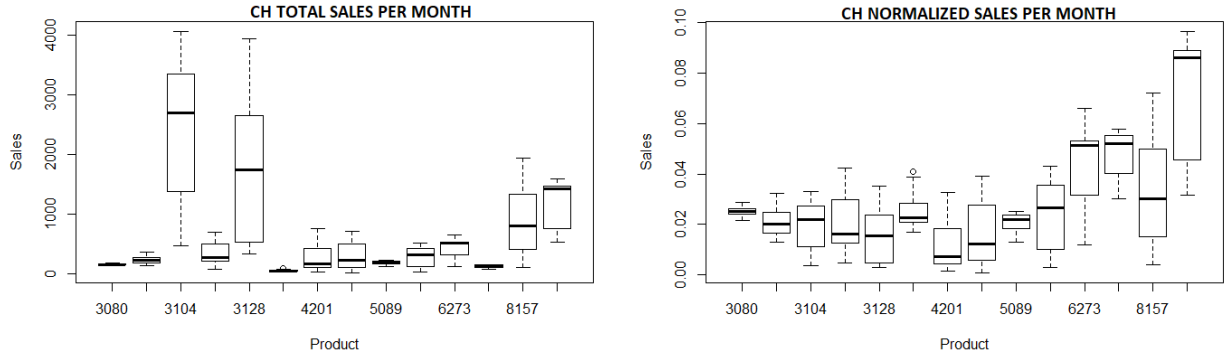
(b) Normalized sales for different products in the US

Figure 5: Comparison of the sales and normalized sales for the different products in the US.

When looking into the China data, Figure 6a, it is possible to observe that there are less variations in the number of sales between products, only products 3104 and 3128 sell more units per month

than the rest. However, it is also possible to see how products sell a smaller number of units in China than in the US. One of the reasons why this happens is because the China market was penetrated after the US one and the competition is higher. At the same time, it is possible to observe the lack of outliers.

After normalizing the China data, Figure 6b, it is possible to see how in general box plots are wider than in the US. This happens mainly because products in China were introduced later than in the US and the number of data points is lower. It is possible to see all the products in detail in Appendix C.



(a) Sales for different products in China

(b) Normalized sales for different products in China

Figure 6: Comparison of the sales and normalized sales for the different products in China.

4.3 Significant differences between markets

The data acquired from the company presents sales and products two different markets: China and the United States. This opens two possibilities of research, either combining sales of both countries and make a general model, or creating different models for the different countries. In order to address which solution is better, a t-test is performed.

A t-test is a statistical hypothesis test used to determine whether the mean of a population significantly differs from a specific value or from the mean of another population. In this situation, a paired sample t-test, also called dependent sample t-test will be used in order to check if the means between two populations differ (China and US).

In order to apply this test to the data, it is first needed to have the same number of observations and same number of products in both countries, as pairs of observations are compared. For that purpose, missing sales values for a specific month are fixed by adding "NA" in the sales variable. The second consideration is to have the same number of products in both countries, in order to do that, products only present in one country are deleted form the data set to perform this test. The products removed from the US market are 3090, 8179 and 8614; whereas in the China market the product 3080 is removed.

Once the same number of observations and products per country are obtained, it is needed to

generated two sorted arrays of data in order to perform the paired t-test. After performing it, the p-value **2.2e-16** is obtained. As the p-value of the test is less than the significance level alpha (0.05), the null hypothesis is rejected and can be concluded that the two populations are significantly different. Therefore, it is not possible to combine the data from the two populations, and models for both markets will need to be generated.

4.4 Clustering

After confirming that both markets will be studied separately, a cluster analysis will be performed for both markets in order to find similarities between products, and with the purpose of develop sets of Bass model parameters for the similar products. This analysis will create a series of clusters in a way that products in the same group are more similar to each other than those products in other groups. The aim of this analysis is to generate different sets of products within a market for the purpose of then generating a forecasting model per cluster, which in theory, should be more accurate than a general model for the whole market. In particular, the k-means clustering method will be applied, which is the most popular method used in data mining.

A step needed to be performed before applying the clustering method is to determine the number of clusters the data will be partitioned in. For that, 3 different methods will be studied: the Elbow method, the Gap statistic method and the Silhouette method. Finally, the Silhouette method will be applied to both markets, as this is the method that generates the smallest number of clusters and not many products are available in each country.

4.4.1 Elbow method

The Elbow method (Zambelli, 2016) looks at the total within-cluster sum of square (WSS), computed as shown in equation (16) (where S_k is the set of observations in the kth cluster and X_{kj} is the jth variable of the cluster center for the kth cluster), as a function of the number of clusters. For that, it computes the k-means method assigning k values from 1 to 10. Then, for each k, it calculates the total within-cluster sum of square (WSS) and plots the curve of WSS according to the number of clusters k. Finally, the location of a bend is considered as the appropriate number of clusters that should be chosen.

$$\sum_{k=1}^{K} \sum_{i \epsilon S_k} \sum_{j=1}^{p} (x_{ij} - \widetilde{x}_{kj})^2 \tag{16}$$

4.4.2 Gap statistic method

The gap statistic method (Tibshirani, Walther, & Hastie, 2001) compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution

of the data. Following the same logic as previous methods, it generates observations for different k values (1 to max), and computes the total within intra-cluster variation for each of those. Then, it repeats this same process but for a second reference data set auto generated with a random uniform distribution. After that, for each k, it computes the estimated gap statistic as the deviation of the observed total within intra-cluster variation for the first data set from its expected value in the second data set, following the formula stated in the equation (17). As the output the smallest value of k such that the gap statistic is within one standard deviation of the gap at $k+1$ should be chosen.

$$Gap_n(k) = E_n\{logW_k\} - logW_k \tag{17}$$

4.4.3 Silhouette method

The Silhouette method (Rousseeuw, 1987) computes the average silhouette metric for all the different observations. This metric is calculated following the formula shown in equation (18), where $a(i)$, represents the average distance between i and the other data within the same cluster, and $b(i)$ represents the smallest average distance of i to all points in any other clusters. It follows the same logic as the Elbow method, using the k-means method and computing this metric for all the k values from 1 to 10. Then a curve showing the average silhouette metric is drawn for all the different k values, where the maximum is considered as the appropriate number of clusters.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{18}$$

When applying this method to the US data, the graph shown in Figure 7 is generated:

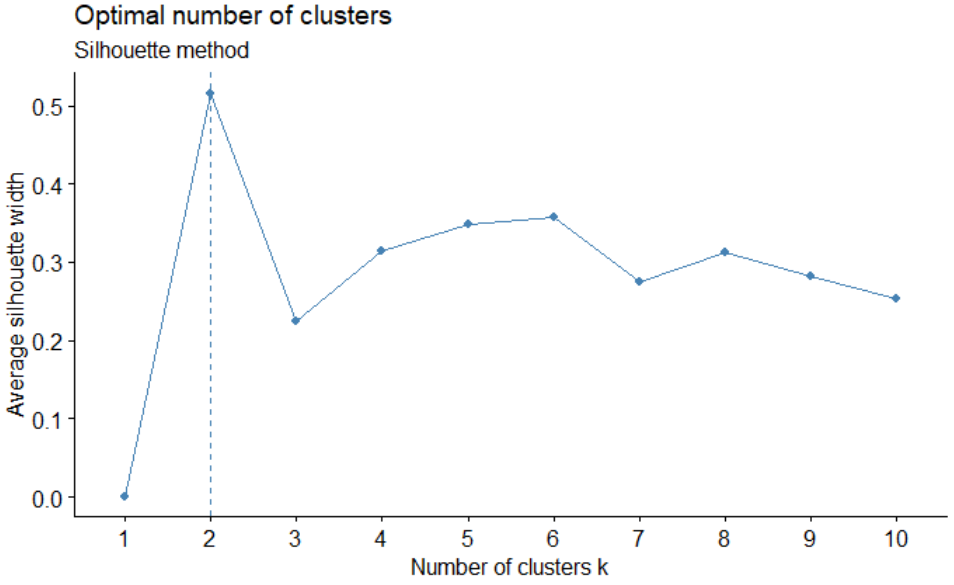


Figure 7: Silhouette method output for the US market.

Therefore, for the US market the optimal number of clusters suggested by the Silhouette method is 2.

Once the number of clusters is selected, the k-means clustering method is applied in R to the US data set. This clustering output is shown in Table 4.

Cluster	Life cycle	Amount	Products
Cluster 1	Ended	4	3090, 4201, 5089, 6273
	Not ended	8	3105, 4395, 6272, 7171, 8157, 8179, 8180, 8614
Cluster 2	Ended	3	3104, 3128, 3159
	Not ended	1	3101

Table 4: Product clusters for the US market

It is possible to observe how 12 products belong to Cluster 1, from which 4 products have finished the life cycle; whereas only 4 products belong to Cluster 2, from which 3 products have finished the life cycle.

When applying the Silhouette method to the China data, the graph shown in 8 is generated:

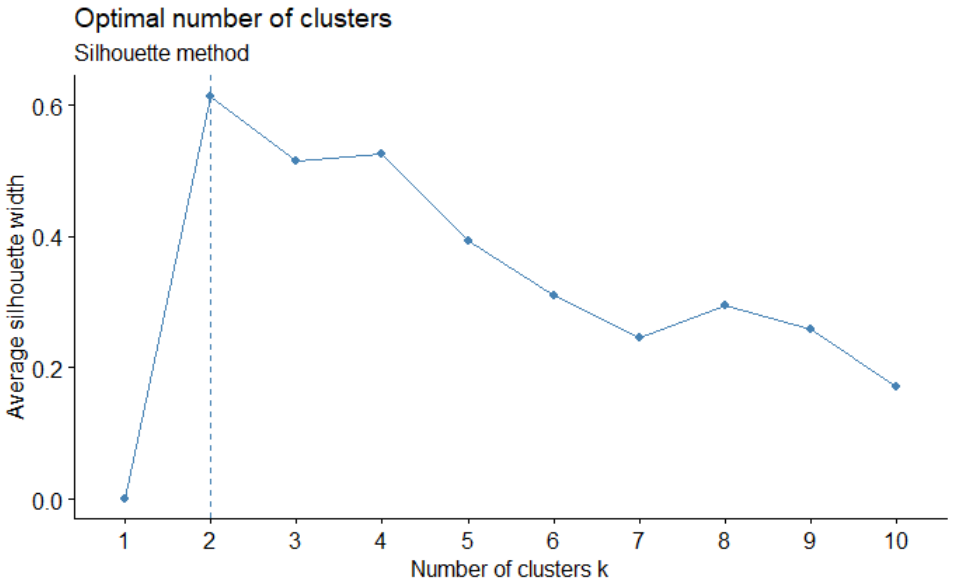


Figure 8: Silhouette method output for the China market.

Therefore, for the China market the optimal number of clusters suggested by the Silhouette method is 2.

Once the number of clusters is selected, and following the same criteria used for the US market, the k-means clustering method is applied in R to the China data set. This clustering output is shown in Table 5.

Cluster	Life cycle	Amount	Products
Cluster 1	Ended	1	4395
	Not ended	10	3101, 3104, 3105, 3128, 5089, 6272, 6273, 7171, 8157, 8180
Cluster 2	Ended	2	3080, 3159
	Not ended	1	4201

Table 5: Product clusters for the China market

It is possible to observe how 11 products belong to Cluster 1, from which only 1 product has finished the life cycle; whereas only 3 products belong to Cluster 2, from which 2 products have finished the life cycle.

4.5 Bass model fitting

The previous introduced Bass forecasting diffusion model can be expressed based on the following equation (19):

$$S(T) = pm + (q - p)Y(T) - (q/m)[Y(T)]^2 \tag{19}$$

Integrating this equation over time, the following differential equation is obtained (20):

$$S(T) = [m(p+q)^2/p]\frac{e^{-(p+q)T}}{[1 + (q/p)e^{-(p+q)T}]^2} \tag{20}$$

Based on those 2 equations, (Massiani & Gohs, 2015) discusses how 3 different methods have been used over the last 50 years to calculate the Bass model parameters based on historic data. The two first methods, maximum likelihood estimation (MLE) and nonlinear least-squares (NLS) techniques are based on the second equation, the differential one. Studies such as the one carried by (Srinivasan & Mason, 1986) show how the NLS approach is slightly better than the MLE regarding accuracy when calculating the parameters. The third method, ordinary least squares (OLS), is based on the first equation, and it is not applicable to the second one, as this second equation is not linear. According to (Massiani & Gohs, 2015), a better performance in calculating parameters from the differential equation (20) rather than from the standard equation (19) has not been found; that is the reason why the OLS method will be used in this research in order to calculate the Bass parameters.

An application of the OLS method to calculate the Bass parameters was defined by (Dodds, 1973). Based on this application, an analogous model from the first equation (21) is generated.

$$S(T) = a + bY(T-1) + c[Y(T-1)]^2 \tag{21}$$

Based on this equation (21), the parameters of the Bass model (q, p and m) can be generated by calculating the regression coefficients (a, b and c), and then solving the following equations

(22)(23)(24):

$$q = -mc \tag{22}$$

$$p = a/m \tag{23}$$

$$m = -\frac{-b - \sqrt{b^2 - 4ac}}{2c} \tag{24}$$

Given the form of the last equation (24), a mathematical solution in real numbers, and therefore an estimate for q, p and m, will exist only if the radicand is not negative. If the radicand is negative, it will not be possible to calculate the values q, p and m, and they will be coded as "NA". Once the three estimations of the parameters are calculated, they will be substituted into the differential equation.

4.5.1 US market regression and parameter generation

Based on the previous methodology, it is possible now to generate estimates for the q, p and m values in both markets (China and US) for all the products; later, in the results chapter, only the values calculated for the products that have already ended the product life cycle will be used to generate the different models. However, at this point, it is interesting to calculate the parameters for all the products to see how accurate those parameters are being generated in the whole population.

In Table 6, the parameters calculated for the US market are presented along with the regression coefficients, the coefficient of determination, R^2, and adjusted coefficient of determination, $Adj.R^2$. The last two coefficients will provide a measure of how well the original sales curve is replicated by the model built from the parameters calculated.

Based on the previous table, it is possible to observe how the accuracy of the parameters calculated is high. As an example, the coefficient of determination is higher than 0.85 for 11 products. However, there are two products, 7171 and 6272 that show a small accuracy in the parameters generated; this is caused because those products are still in a long maturity phase, where the sales curve is a straight line, and therefore the quadratic regression does not fit properly the data points. It is also possible to see how the average value for the parameter p is around 0.015 whereas the average value for the parameter q is 0.1. The parameter m varies more from one product to another as explained in previous chapters different products focus on different markets segments and therefore are expected to present different numbers of sales. Finally, there are three products (4395, 6272, and 7171) for which it was not possible to calculate parameters p, q and m due to a negative radicand. Therefore, those products will not be used in the performance analysis chapter during the training of the model.

Product	End PLC	a	b	c	R^2	Adj. R^2	p	q	m
3090	Yes	214.64	0.1044	-5.5 e-06	0.9075	0.9025	0.1030	0.1147	20836
3101	No	-937.78	0.0439	-7.4 e-08	0.8636	0.8583	-0.0016	0.0422	569801
3104	Yes	3479	0.0336	-1.7 e-07	0.8894	0.8855	0.0133	0.0469	261183
3105	No	-176.05	0.0473	-8.3e-08	0.8822	0.8774	-0.0018	0.0454	94496
3128	Yes	1085.5	0.0338	-8.3e-08	0.9157	0.9133	0.0025	0.0363	433851
3159	Yes	4064.3	0.0428	-1.3e-07	0.6913	0.6782	0.0105	0.0534	384234
4201	Yes	605.4	0.0680	-5.7e-07	0.8875	0.8830	0.0048	0.0728	125917
4395	No	104.1	-0.06	1.8e-05	0.8901	0.8830	NA	NA	NA
5089	Yes	1608.9	0.0199	-2.9e-07	0.4215	0.3933	0.0139	0.0339	115128
6272	No	6833.5	0.002	2.1e-09	0.2197	0.1817	NA	NA	NA
6273	Yes	1409.8	0.0229	-3.5e-07	0.5885	0.5650	0.01374	0.0366	102538
7171	No	1701.2	-0.0215	4.2e-07	0.1823	0.1296	NA	NA	NA
8157	No	171.56	0.0952	-3.0e-06	0.9875	0.9865	0.0052	0.1004	32524
8179	No	1718.9	0.0042	-7.8e-07	0.9438	0.9382	0.0347	0.0390	49425
8180	No	1174.1	0.2388	-4.3e-06	0.9217	0.9044	0.0199	0.2587	58825
8614	No	1229.5	0.3822	-6.6e-06	0.9729	0.9549	0.0202	0.4025	60748

Table 6: Bass model parameters calculated for the US market

4.5.2 China market regression and parameter generation

The same methodology is applied to the China products. Table 7 shows the output of this process.

Looking at the table, it is possible to observe how as it already happened in the US market, the accuracy of the parameters calculated is high. As an example, the coefficient of determination is higher than 0.85 for 11 products. Only the product 3080 is showing a low accuracy in the parameter generation. This is caused because of the missing data this product is presenting at the beginning of the life cycle. Regarding the average p value, in this situation it is 0.005 whereas the average q value is 0.082. Both values are slightly smaller than the ones generated for the US market. In this case, there are two products (3101 and 3159) for which it was not possible to calculate the parameter p, q and m.

Product	End PLC	a	b	c	R^2	Adj. R^2	p	q	m
3080	Yes	150.71	0.0043	-6.8e-07	0.0373	-0.0161	-0.0081	0.0125	18387
3101	No	177.51	-0.0005	1.4e-06	0.7226	0.7105	NA	NA	NA
3104	No	661.81	0.0592	-2.7e-07	0.9770	0.9761	0.0029	0.0621	225139
3105	No	107.86	0.0380	7.6e-08	0.9721	0.9709	-0.0002	0.0378	494615
3128	No	158.41	0.0690	-3.6e-07	0.9525	0.9509	0.0008	0.0698	191804
3159	Yes	85.88	-0.0521	1.4e-05	0.9172	0.9128	NA	NA	NA
4201	No	41.28	0.0346	-2.3e-07	0.9768	0.9762	0.0002	0.0349	148170
4395	Yes	43.15	0.1315	-6.4e-06	0.9134	0.9104	0.0021	0.1294	20153
5089	No	136.04	0.0376	-4.1e-06	0.8218	0.8137	0.0115	0.0491	11814
6272	No	53.33	0.0965	-5.5e-06	0.9732	0.9718	0.0029	0.0994	17973
6273	No	170.19	0.1043	-6.6e-06	0.8914	0.8806	0.0099	0.1142	17188
7171	No	71.244	0.0729	-1.7e-05	0.9859	0.9844	0.0140	0.0869	5075
8157	No	186.59	0.0983	-1.3e-06	0.9919	0.9913	0.0024	0.1007	77045
8180	No	408.84	0.1866	-7.4e-06	0.9122	0.8963	0.0151	0.2018	27006

Table 7: Bass model parameters calculated for the China market

4.5.3 Total Sales Regression and parameter generation

Finally, and following the previous methodology, Bass model parameters are generated after computing the total number of sales in the US, China, and in both countries together (Global Total Sales). Table 8 shows the results of this process:

Market	a	b	c	R^2	Adj. R^2	p	q	m
US Total Sales	14663	0.0546	-4.4e-08	0.8371	0.8312	0.0100	0.0647	1458947
CH Total Sales	111.58	0.0674	-2.2e-06	0.6338	0.6209	0.0035	0.0709	31089
Global Total Sales	14646	0.0557	-4.4e-08	0.8340	0.8281	0.0099	0.0656	1477913

Table 8: Bass model parameters calculated for the total sales of the US and China markets

From the table, it is possible to observe how the accuracy of the results is high for the US and Global market, whereas for the China market a lower accuracy of 0.63 is achieved. It is also possible to see how the accuracy and parameters in the US and Global markets are close to each other; this is caused because more data points and a higher number of sales is present in the US market than in the China market.

Finally, and looking at the tables 6, 7 and 8, it possible to see how a better fitting is acquired when the Bass model parameters are generated from an individual product, rather than from the total sales of all of them together.

5 Results

The purpose of this chapter is to build different forecasting models based on the parameters acquired in the previous section and the equations generated from the Bass model, and finally, validate them against the literature and company estimates. In total, 10 different models have been generated, 3 focusing in the US market, 3 focusing in the China market, 1 global model averaging the US and China models, and, finally, 3 models based on the total sales regression for the US and China markets.

5.1 US market

For the US market 3 sales forecasting models have been generated: one from all the products that have finished the life cycle, one from the previous generated cluster 1 and once from the cluster 2. The first model is generated by averaging the p and q values acquired for all the products that have already finished the life cycle in this market (p: 0.0098, q: 0.0564). The second model focus on the cluster 1 previously generated, and it is based on the averaged p and q values acquired for the products which have finished their life cycle within this cluster (p: 0.0107, q: 0.0645). Finally, the third model focus on the averaged p and q values for the products that have finished their life cycle in the second cluster (p: 0.0088, q: 0.0455). Once the three models are generated, they are tested again those products which have not finished yet their life cycles within this market, by calculating the statistical measure R-squared which computes how close the data is to the fitted regression line. It is possible to observe the three models' parameters and the coefficient of determination obtained for each of them in Table 9. Furthermore, Table 10 shows the averaged coefficient of determination for each of the models.

	US Averaged Model (M1)				**US Averaged Cluster 1 (M2)**				**US Averaged Cluster 2 (M3)**			
Product	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**
3101	0.0098	0.0564	1	0.1940	X	X	X	X	0.0088	0.0455	1	0.6488
3105	0.0098	0.0564	1	0.7292	0.0107	0.0645	1	0.7838	X	X	X	X
8157	0.0098	0.0564	1	0.9479	0.0107	0.0645	1	0.9193	X	X	X	X
8179	0.0098	0.0564	1	0.8788	0.0107	0.0645	1	0.8644	X	X	X	X
8180	0.0098	0.0564	1	0.6632	0.0107	0.0645	1	0.6637	X	X	X	X
8614	0.0098	0.0564	1	0.9427	0.0107	0.0645	1	0.9424	X	X	X	X

Table 9: Life cycle models for the US market

Model	**p**	**q**	**Averaged R^2**
US Averaged Model (M1)	0.0098	0.0564	0.7259
US Averaged Cluster 1 Model (M2)	0.0107	0.0645	0.8347
US Averaged Cluster 2 Model (M3)	0.0088	0.0455	0.6488

Table 10: Averaged coefficients of determination for the US market

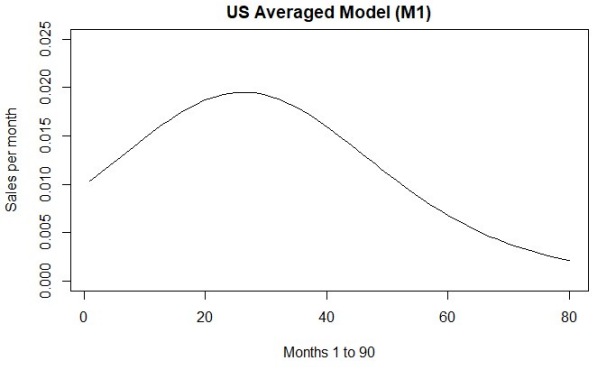
Based on the two previous tables, it is possible to observe how the best accuracy is abstained by the model generated for the cluster number 1. It is also possible to see how the models present a good fit for some specific products, such as model 1 for products 8157 (0.9479) and 8614 (0.9427), at the same time the same model presents bad estimates for a different product 3101 (0.1940). Therefore, it suggests that fitting the same model to different products may be useful, but the model will not always provide a high accuracy.

Estimates of the number of months involved in the product life cycle are also calculated, and they are shown in Table 11, where it is possible to observe how cluster 2 presents a considerably longer product life cycle that cluster 1.

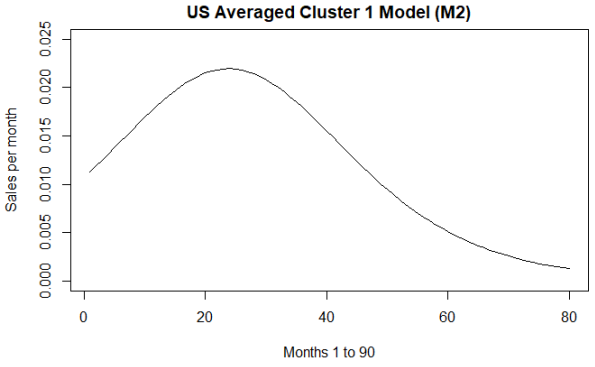
	US Averaged Model (M1)	**US Averaged Cluster 1 (M2)**	**US Averaged Cluster 2 (M3)**
Life cycle	61	55	73

Table 11: Product life cycle (in months) per model in the US market

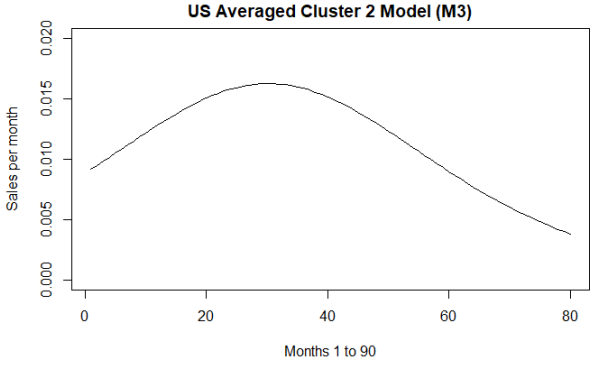
Finally, curves for all the three different models are shown in Figure 9, and it is possible to see the particular fit of each curve for each product in Appendix D.



(a) US Market Model (M1)



(b) US Market Cluster 1 Model (M2)



(c) US Market Cluster 2 Model (M3)

Figure 9: Product life cycle curves per model in the US market.

5.2 China market

In the China market, the same methodology than in the US market is applied, and 3 different models are generated: for all products, for cluster 1 and for cluster 2. The first one is obtained by averaging the p and values for all the products which have finished the life cycle in the market (p: 0.0030, q: 0.0710). The second model focus on cluster 1, generating the following p and q parameters (p: 0.002, q: 0.1294). In this situation, it was not possible to create the third model based on the second cluster; this is caused because a negative p value was generated, and as explained by (Orbach, 2016) it should be discarded. After generating both models, those are tested against the products that have not finished the product life cycle yet. It is possible to observe those results in Table 12, and the averaged coefficient of determination for both models in Table 13.

	CH Averaged Model (M4)				**CH Averaged Cluster 1 (M5)**				**CH Averaged Cluster 2 (M6)**			
Product	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**
3104	0.0030	0.0710	1	0.7130	0.0021	0.1294	1	0.3393	X	X	X	X
3105	0.0030	0.0710	1	0.7996	0.0021	0.1294	1	0.1051	X	X	X	X
3128	0.0030	0.0710	1	0.4724	0.0021	0.1294	1	0.0374	X	X	X	X
4201	0.0030	0.0710	1	0.1600	X	X	1	X	-0.0081	0.0125	NA	NA
5089	0.0030	0.0710	1	0.1233	0.0021	0.1294	1	0.2217	X	X	X	X
6272	0.0030	0.0710	1	0.9720	0.0021	0.1294	1	0.9381	X	X	X	X
6273	0.0030	0.0710	1	0.7982	0.0021	0.1294	1	0.6820	X	X	X	X
7171	0.0030	0.0710	1	0.8507	0.0021	0.1294	1	0.7281	X	X	X	X
8157	0.0030	0.0710	1	0.9797	0.0021	0.1294	1	0.9864	X	X	X	X
8180	0.0030	0.0710	1	0.7656	0.0021	0.1294	1	0.6794	X	X	X	X

Table 12: Life cycle models for the China market

Model	**p**	**q**	**Averaged R^2**
CH Averaged Model (M4)	0.0030	0.0710	0.6634
CH Averaged Cluster 1 Model (M5)	0.0021	0.1294	0.5241
Averaged Cluster 2 Model (M6)	NA	NA	NA

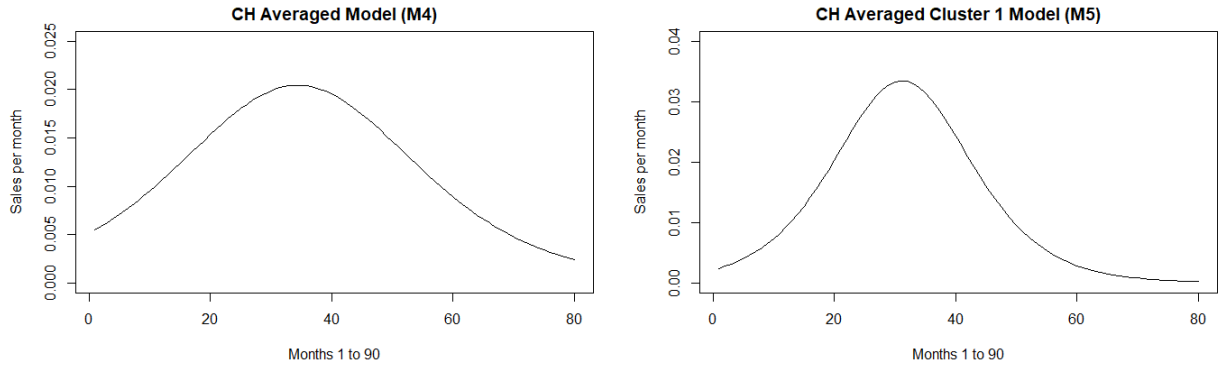
Table 13: Averaged coefficients of determination for the China market

It is possible to see how for the China market the averaged model for the whole country is getting a better coefficient of determination than the model for the cluster 1. At the same time, it is possible to observe how the models for the US market obtained a better fit; and also, as already happened in the US market, how some models show a high accuracy for certain products (model 4 and product 8157) at the same time they show a low accuracy for other products (model 4 and product 5089).

Regarding the estimates of the number of months involved in the product life cycle, which can be observed in Table 14, it is possible to see how the model for the cluster 1 shows a smaller number of months in the product life cycle than the general model. Finally, Figure 10 shows the different curves generated for the different models, and it is possible to see the particular fit of each curve for each product in Appendix E.

	CH Averaged Model (M4)	CH Averaged Cluster 1 (M5)	CH Averaged Cluster 2 (M6)
Life cycle	73	48	NA

Table 14: Product life cycle (in months) per model in the China market



(a) China Market Model (M4)

(b) China Market Cluster 1 Model (M5)

Figure 10: Product life cycle curves per model in the China market.

5.3 Global market model

The purpose of this section is to study how the US Averaged Model (M1) and the China Averaged Model (M4) perform in both markets, and to generate a new Global Averaged Model (M7) which is obtained by averaging the parameters p and q that were generated for the US and China averaged models (M1 and M4). Those three models are compared as their parameter have been generated by using all the products that have finished their life cycle in the different countries. After generating this new global model, all models are tested against the products that have not finished the product life cycle yet in both markets. It is possible to observe the results in Table 15, and the averaged coefficient of determination for all three models in both markets in Table 16.

		US Averaged Model (M1)				**CH Averaged (M4)**				**Global Averaged (M7)**			
Country	**Product**	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**
US	3101	0.0098	0.0564	1	0.1940	0.0030	0.0710	1	0.7742	0.0083	0.0596	1	0.4970
US	3105	0.0098	0.0564	1	0.7292	0.0030	0.0710	1	0.4074	0.0083	0.0596	1	0.7488
US	8157	0.0098	0.0564	1	0.9479	0.0030	0.0710	1	0.9714	0.0083	0.0596	1	0.9690
US	8179	0.0098	0.0564	1	0.8788	0.0030	0.0710	1	0.9256	0.0083	0.0596	1	0.9005
US	8180	0.0098	0.0564	1	0.6632	0.0030	0.0710	1	0.6064	0.0083	0.0596	1	0.6458
US	8614	0.0098	0.0564	1	0.9427	0.0030	0.0710	1	0.9375	0.0083	0.0596	1	0.9413
CH	3104	0.0098	0.0564	1	0.0241	0.0030	0.0710	1	0.7130	0.0083	0.0596	1	0.2193
CH	3105	0.0098	0.0564	1	0.0100	0.0030	0.0710	1	0.7996	0.0083	0.0596	1	0.0476
CH	3128	0.0098	0.0564	1	0.3857	0.0030	0.0710	1	0.4724	0.0083	0.0596	1	0.1840
CH	4201	0.0098	0.0564	1	0.8233	0.0030	0.0710	1	0.1600	0.0083	0.0596	1	0.7667
CH	5089	0.0098	0.0564	1	0.7065	0.0030	0.0710	1	0.1233	0.0083	0.0596	1	0.4798
CH	6272	0.0098	0.0564	1	0.6508	0.0030	0.0710	1	0.9720	0.0083	0.0596	1	0.8435
CH	6273	0.0098	0.0564	1	0.8798	0.0030	0.0710	1	0.7982	0.0083	0.0596	1	0.8562
CH	7171	0.0098	0.0564	1	0.9305	0.0030	0.0710	1	0.8507	0.0083	0.0596	1	0.9075
CH	8157	0.0098	0.0564	1	0.8530	0.0030	0.0710	1	0.9797	0.0083	0.0596	1	0.9137
CH	8180	0.0098	0.0564	1	0.8139	0.0030	0.0710	1	0.7656	0.0083	0.0596	1	0.7996

Table 15: Life cycle models for the Global market

Country	**Model**	**p**	**q**	**Averaged R^2**
US	US Averaged Model (M1)	0.0098	0.0564	0.7259
US	CH Averaged Model (M4)	0.0030	0.0710	0.7704
US	Global Averaged Model (M7)	0.0083	0.0596	0.7837
CH	US Averaged Model (M1)	0.0098	0.0564	0.6077
CH	CH Averaged Model (M4)	0.0030	0.0710	0.6634
CH	Global Averaged Model (M7)	0.0083	0.0596	0.6017

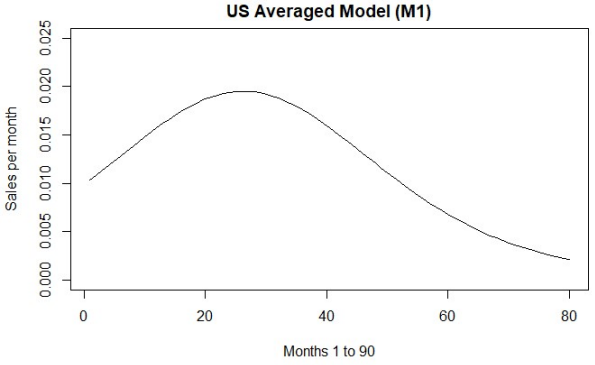
Table 16: Averaged coefficients of determination for the Global market

Based in the previous tables it is possible to see how the Global Averaged Model is the one getting the best performance in the US, whereas the China Averaged Model is the one getting the best results in the China market. It is also surprising to see how the China Averaged model is getting a better performance in the US market than the US Averaged Model. This reinforces insights from the previous sections that spot how the same model has different performance on different products. This suggests that models should be based on a product level, trying to identify similar products; rather than on a market level, where a model is developed to be applied to different products, and therefore, the general accuracy is lowered.

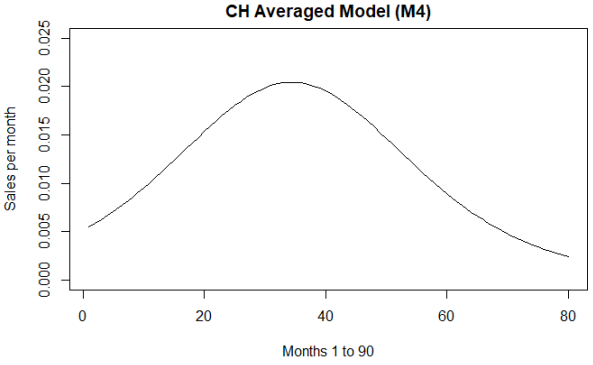
The number of months involved in the product life cycle for each model are shown in Table 17. It is possible to observe how the US Averaged Model and the Global Averaged Model are close to each other, mainly because the number of data points is bigger in the US market than in the China market. Finally, Figure 11 shows the different curves generated for the different models. It is possible to observe how in the China market products need more time to penetrate than in the US market. Appendix F shows the particular fit of each curve for each product.

	US Averaged Model (M1)	**CH Averaged Model (M4)**	**Global Averaged Model (M7)**
Life cycle	61	73	63

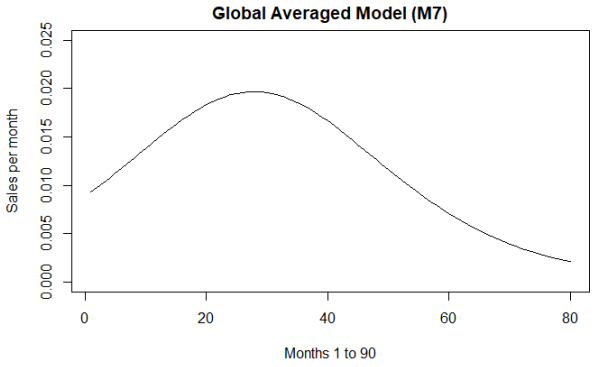
Table 17: Product life cycle (in months) per model in the Global market



(a) US Market Model (M1)



(b) China Market Model (M4)



(c) Global Market Model (M7)

Figure 11: Product life cycle curves per model in the Global market.

5.4 Total Sales per Market Model

Following the previous methodology, three models are generated for the parameters generated after combining the total sales in the US and China market in the section 4.5.3. The first one is generated for the US market by using the parameters (p: 0.0100, q: 0.0647). The second one is generated for the China market using the following parameters (p: 0.0035, q: 0.0709). The third model is based on the combination of the total sales in the US and China markets together (p: 0.0099, q: 0.0656). The three models generated are tested by calculating the coefficient of determination in both countries, US and China. This can be observed in Table 18. Furthermore, Table 19 shows the averaged coefficient of determination for all three models in both markets, as well as the averaged coefficient of determination for the models presented in section 5.3 (M1, M4 and M7).

		US Total Sales (M8)				**CH Total Sales (M9)**				**Global Total Sales (M10)**			
Country	**Product**	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**	**p**	**q**	**m**	**R^2**
US	3101	0.0100	0.0647	1	0.8184	0.0035	0.0709	1	0.7535	0.0099	0.0656	1	0.8228
US	3105	0.0100	0.0647	1	0.7683	0.0035	0.0709	1	0.1190	0.0099	0.0656	1	0.7669
US	8157	0.0100	0.0647	1	0.9396	0.0035	0.0709	1	0.9528	0.0099	0.0656	1	0.9425
US	8179	0.0100	0.0647	1	0.8775	0.0035	0.0709	1	0.9271	0.0099	0.0656	1	0.8801
US	8180	0.0100	0.0647	1	0.6570	0.0035	0.0709	1	0.5933	0.0099	0.0656	1	0.6548
US	8614	0.0100	0.0647	1	0.9419	0.0035	0.0709	1	0.9363	0.0099	0.0656	1	0.9417
CH	3104	0.0100	0.0647	1	0.0002	0.0035	0.0709	1	0.9349	0.0099	0.0656	1	3.0e-07
CH	3105	0.0100	0.0647	1	0.0773	0.0035	0.0709	1	0.7143	0.0099	0.0656	1	0.0773
CH	3128	0.0100	0.0647	1	0.4753	0.0035	0.0709	1	0.2898	0.0099	0.0656	1	0.4655
CH	4201	0.0100	0.0647	1	0.7981	0.0035	0.0709	1	0.2771	0.0099	0.0656	1	0.7931
CH	5089	0.0100	0.0647	1	0.7793	0.0035	0.0709	1	0.0139	0.0099	0.0656	1	0.7735
CH	6272	0.0100	0.0647	1	0.4942	0.0035	0.0709	1	0.9449	0.0099	0.0656	1	0.5082
CH	6273	0.0100	0.0647	1	0.8804	0.0035	0.0709	1	0.7700	0.0099	0.0656	1	0.8782
CH	7171	0.0100	0.0647	1	0.9322	0.0035	0.0709	1	0.8227	0.0099	0.0656	1	0.9301
CH	8157	0.0100	0.0647	1	0.8183	0.0035	0.0709	1	0.9936	0.0099	0.0656	1	0.8238
CH	8180	0.0100	0.0647	1	0.8104	0.0035	0.0709	1	0.7528	0.0099	0.0656	1	0.8087

Table 18: Life cycle models for the Total Sales per market

Based on the previous tables it is possible to observe how between the new models, the US Total Sales Model (M8) is the one presenting a better accuracy in the US market, whereas the Global Total Sales Model (M10) is the on presenting the best accuracy in the China market. (Note that product 3104 in the China market was not considered when computing the averaged R^2 in the model M10, as the coefficient of determination shows a really small value for that product, not being in line with the rest of the products, and therefore it is considered as an outlier).

It is also possible to see how the accuracy of the Total Sales Models is better than the one presented by the previous models (M1, M4, M7). This suggests that adding the total number of sales for all the products in a market, and then calculate the Bass model parameters is a better strategy than generating the Bass model parameters per product and then averaging them. Furthermore, and as it happened in previous sections, it is possible to observe how the Total Sales Models present different levels of accuracy depending on the product, meaning than a one on one forecasting model is preferred rather than a global forecasting model.

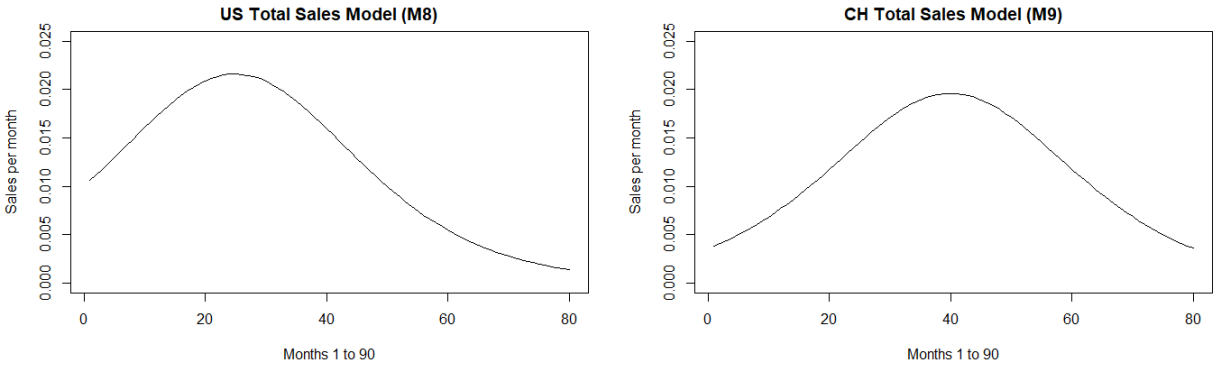
Country	Model	p	q	Averaged R^2
US	US Averaged Model (M1)	0.0098	0.0564	0.7259
US	CH Averaged Model (M4)	0.0030	0.0710	0.7704
US	Global Averaged Model (M7)	0.0083	0.0596	0.7837
US	US Total Sales Model (M8)	0.0100	0.0647	0.8337
US	CH Total Sales Model (M9)	0.0035	0.0709	0.7136
US	Global Total Sales Model (M10)	0.0099	0.0656	0.8348
CH	US Averaged Model (M1)	0.0098	0.0564	0.6077
CH	CH Averaged Model (M4)	0.0030	0.0710	0.6634
CH	Global Averaged Model (M7)	0.0083	0.0596	0.6017
CH	US Total Sales Model (M8)	0.0100	0.0647	0.6065
CH	CH Total Sales Model (M9)	0.0035	0.0709	0.6514
CH	Global Total Sales Model (M10)	0.0099	0.0656	0.6731*

Table 19: Averaged coefficients of determination for the US, China and Total Sales models

Table 20 shows the estimates of the number of months involved in the product life cycle. It is possible to see how the model generated for the China market present a bigger number of months whereas the US and Global model are similar to each other. Finally, Figure 12 shows the different curves generated for the three different models, where it is possible to see how the penetration is slower in the China market as explained in previous sections. Appendix G shows the particular fit of each curve for each product.

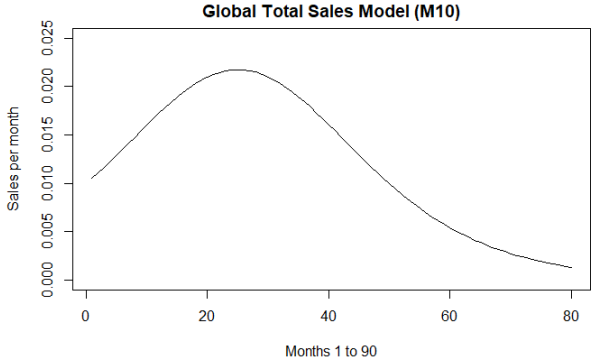
	US Total Sales (M8)	CH Total Sales (M9)	Global Total Sales (M10)
Life cycle	56	70	55

Table 20: Product life cycle (in months) per model in the Total Sales models



(a) US Total Sales Model (M8)

(b) China Total Sales Model (M9)



(c) Global Total Sales Model (M10)

Figure 12: Product life cycle curves per model for the Total Sales Models.

5.5 Cross Validation

The results are validated by looking into the literature which parameters have been obtained by different authors for similar industries; and also, by looking into company estimates for the different products.

5.5.1 Literature estimates

The first step taken in order to compare the results with literature estimates is try to find other studies where Bass model parameters for electric toothbrushes are generated. Unfortunately, and in line with the insights provided in the problem statement chapter, suggesting that small research has been performed in the applicability of diffusion models to the health tech industry, only one estimate for electric toothbrushes has been found. In this paper (Kim & Bass, 2005) the estimates were generated by using two different methods, the already introduced NLS method for which p: 0.1021 and q: 0.483 values were obtained; and the Virtual Bass Model (VBM) method (Zhengrui, Bass, & Bass, 2005) for which p: 0.0617 and q: 0.5235 values were obtained.

When comparing those values with the values estimated in this research, Table 17, it is possible to observe how the values retrieved from the literature are bigger. This is caused because in the Kim and Frank paper, the Bass model is applied on a yearly basis whereas in this research it is applied on a monthly basis, getting therefore more data points, what leads to a reduction in the value of the innovation and imitation coefficients. However, when looking at the ratio between parameters, it is possible to observe how the ratio q to p, in the paper retrieved is 4.73:1 for the NLS method and 8.48:1 for the VBM; whereas in this research, after averaging the results in Table 17, an 8.73:1 ratio is acquired. Therefore, it is possible to confirm that the acquired estimates are in line with the only research performed on electric toothbrushes.

Apart from specific parameters estimations for electric toothbrushes, it is also interesting to study how the parameters varies across different countries, as two markets were studied in this research (US and China). For those specific markets it was studied (den Bulte., 2005) how, in general, the coefficient of innovation (p) in Asia is half the one in the US, and how the coefficient of imitation (q) in Asia is a quarter less than the one in the US. On the one hand, looking at Table 17, is possible to see how the coefficient of innovation (p) for the US Averaged model is 0.0098, whereas it is 0.0030 for the CH Averaged model. Therefore, in this research, the p coefficient in China is three times smaller than in the US, what is close to the previously cited study. On the other hand, the coefficient of imitation (q) for the US averaged model is 0.0564, whereas it is 0.0710 for the CH averaged model. Therefore, in this situation the result is not in line with the cited paper, being the q parameter slightly bigger in China than in the US, and not a quarter less as proposed by the paper.

5.5.2 Company estimates

Aiming to validate the results generated in the previous sections from a company point of view, a survey was sent to different managers within the organization in order to obtain the life cycle estimates, generated by the company, for the different products that have not finished the product life cycle yet. The structure of the survey can be observed in Appendix B. After sending the survey, three managers answered it providing estimates for the different products. Those estimates were averaged and can be found in the column "Survey" in the Table 21. This table also presents the estimates generated by the different models.

Country	Product	Survey	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
US	3101	144	61	X	55	73	X	X	63	56	70	55
US	3105	120	61	73	X	73	X	X	63	56	70	55
US	8157	42	61	73	X	73	X	X	63	56	70	55
US	8179	48	61	73	X	73	X	X	63	56	70	55
US	8180	48	61	73	X	73	X	X	63	56	70	55
US	8614	48	61	73	X	73	X	X	63	56	70	55
US Avg.	All	75	61	73	X	73	X	X	63	56	70	55
CH	3104	120	61	X	X	73	48	X	63	56	70	55
CH	3105	108	61	X	X	73	48	X	63	56	70	55
CH	3128	60	61	X	X	73	48	X	63	56	70	55
CH	4201	84	61	X	X	73	X	NA	63	56	70	55
CH	5089	72	61	X	X	73	48	X	63	56	70	55
CH	6272	78	61	X	X	73	48	X	63	56	70	55
CH	6273	64	61	X	X	73	48	X	63	56	70	55
CH	7171	60	61	X	X	73	48	X	63	56	70	55
CH	8157	48	61	X	X	73	48	X	63	56	70	55
CH	8180	60	61	X	X	73	48	X	63	56	70	55
CH Avg.	All	75.4	61	X	X	73	48	X	63	56	70	55

Table 21: Product life cycle (in months) per model in the different markets plus manager estimations

From the previous table it is possible to observe how two products in the US market (3101 and 3105) and two products in the China market (3104 and 3105) present a long life cycle estimates, having been in the market for more than 90 months. It is also possible to see how the average estimates for the US (75 months) and the China market (75.4) are similar. Based on those company averaged estimates, model US Averaged Cluster 1 Model (M2) for the US market, and models CH Averaged Model (M4) and CH Total Sales Model (M9) for both markets present the best estimates.

However, looking at each product one by one, it is possible to observe that those estimates are no longer accurate, as for example product 8157 in the US presents a small estimate (42 months) whereas product 3102 in China presents a big estimate (120 months). Therefore, the best way of forecasting the number of months seems to be the applicability of the Bass model one on one: finding a product which has already finished the life cycle, similar to the product for which we want to estimate sales, and generate the model parameters from this first product.

6 Conclusion

The aim of this chapter is to present the conclusions generated from this research. After that, the research limitations will be mentioned, and, finally, future study suggestions will be provided.

6.1 Conclusions

A series of conclusions can be drawn from this research:

- It is possible to apply the Bass forecasting model to a Health Tech company and generate its parameters from historical sales with high accuracy.
- The two markets studied, US and China, present significant differences, being the number of sales per month higher in the US market; whereas the China market presents a longer penetration phase.
- Business insights must be taken into consideration before applying the Bass forecasting model, such as the artificial extension of the life cycle presented in the section 4.1, otherwise, the accuracy of the parameters generated will be small.
- When trying to generate a general Bass forecasting model for all the products in one data set, it is preferred to first generate the total number of sales for all the products within that data set, rather than generate the Bass parameters per product and then averaging them, as it increases the general accuracy.
- Generating a general Bass forecasting model for several products may be useful in those situations where a lack of information is present and some estimates are required, or high accuracy is not needed. However, if high accuracy is needed, it is better to apply the Bass forecasting model on a one on one basis: finding a product which has already finished the life cycle, similar to the product for which we want to estimate sales, and generate the model parameters from this first product.

6.2 Research limitations

One of the main limitations of this research is the lack of data for some products. There is only data available in the company systems from the year 2011 onwards; however, some products were introduced into the market before that date. Having access to that missing data would increase the number of available products and the accuracy of the estimations for some products that currently do not have data for the complete product life cycle.

A second limitation, also related with the previous one, is the number of total products. There are 30 available products in total, from which only 10 have finished the whole life cycle. Due to this fact, and among the different clustering recommendations provided in chapter 4.2, the final number

of clusters to be selected per market had to be 2. Increasing the number of clusters would mean to increase the quality of the estimations, as estimations would be performed closer to the one on one approach (one reference product for each product that wants to be forecasted). However, in the current situation, increasing the number of clusters would mean that some clusters would not present any product that has finished the product life cycle.

Finally, only data from two different markets is available. By increasing the number of markets available, it would be possible to clarify which level of accuracy can be obtained by generating a model for one country and extrapolate it to other countries.

6.3 Future study

The focus of this study was to study and to introduce a sales forecasting model to a health tech company. Once the Bass forecasting model has been proved to be a good model for forecasting sales and for producing estimates in such a company, the immediate next step seems to be the applicability of one or more models suggested in the chapter 2 to the same company, performing an accuracy comparison between them.

Another line of research can be defined by extrapolating the applicability of the model to other products within the company, and to different markets as explained in section 6.2. This way it could be confirmed if this is a robust model for the whole industry and not only for one specific product in two specific markets.

Finally, it would be interesting to link the model parameters, such as the coefficient of innovation p and the coefficient of imitation q, to different marketing strategies within the company; trying to understand how marketing strategies affect the number of innovators and imitators that purchase the different products.

A Variables description

Following the 9 variables used in the final data set are described.

1. **Product**: numeric variable. Defines a specific product.
2. **Month**: numeric variable (1-89). Defines the specific month.
3. **Country**: text variable [US – CH]. Defines the country United States (US) or China (CH).
4. **Sales**: numeric variable. Number of sales per month.
5. **Cumulative Sales**: numeric variable. Amount of sales of a product up to that month.
6. **Squared Cumulative Sales**: numeric variable. Squared amount of sales of a product up to that month.
7. **Normalized Sales**: numeric variable (0 – 1). Normalized number of sales per month.
8. **Cumulative Normalized Sales**: numeric variable. Normalized amount of sales of a product up to that month.
9. **Squared Cumulative Normalized Sales**: numeric variable. Squared normalized amount of sales of a product up to that month.

B Survey

Product life cycle model Survey

This survey is part of my Master thesis in which I studied the applicability of a sales forecasting model to Philips toothbrushes products. The aim of this survey is to validate the results generated in the thesis, focusing in the expected life cycle of different toothbrushes. The estimated answering time is less than 5 minutes, being all the answers to be provided numeric. The answers of this survey will be anonymous.

Name (It will become anonymous)

Role (It will become anonymous)

US Market

What is the expected product life cycle for the product 3101 "Confidential description" in the US market (in number of months e.g. 50)

What is the expected product life cycle for the product 3105 "Confidential description" in the US market (in number of months e.g. 50)

What is the expected product life cycle for the product 8157 "Confidential description" in the US market (in number of months e.g. 50)

What is the expected product life cycle for the product 8179 "Confidential description" in the US market (in number of months e.g. 50)

What is the expected product life cycle for the product 8180 "Confidential description" in the US market (in number of months e.g. 50)

What is the expected product life cycle for the product 8614 "Confidential description" in the US market (in number of months e.g. 50)

China Market

What is the expected product life cycle for the product 3104 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 3105 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 3128 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 4201 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 5089 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 6272 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 6273 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 7171 "Confidential description" in the China market (in number of months e.g. 50)

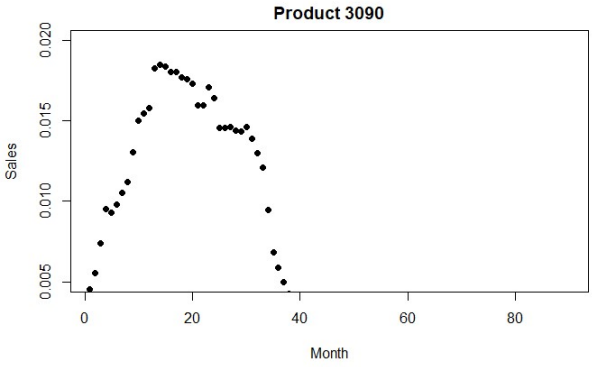
What is the expected product life cycle for the product 8157 "Confidential description" in the China market (in number of months e.g. 50)

What is the expected product life cycle for the product 8180 "Confidential description" in the China market (in number of months e.g. 50)

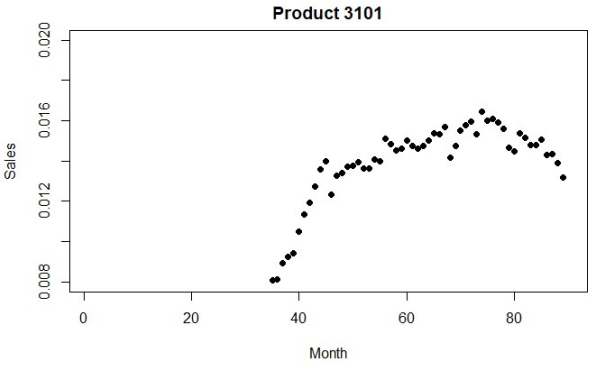
C Products

This section shows the product data that has been used in the research.

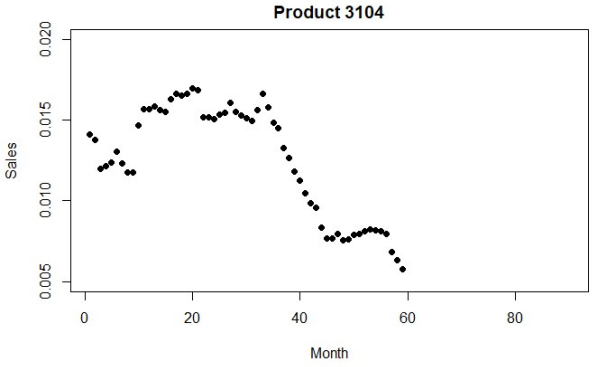
C.1 US products



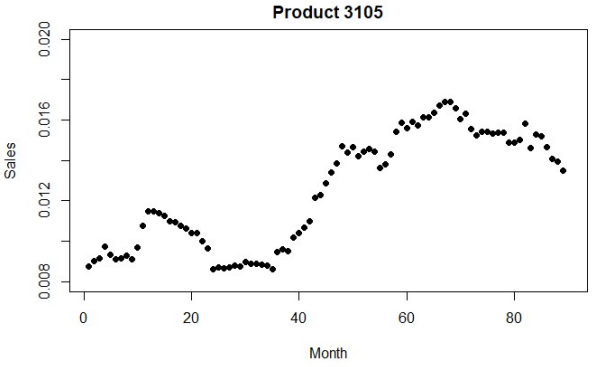
(a) US Product 3090



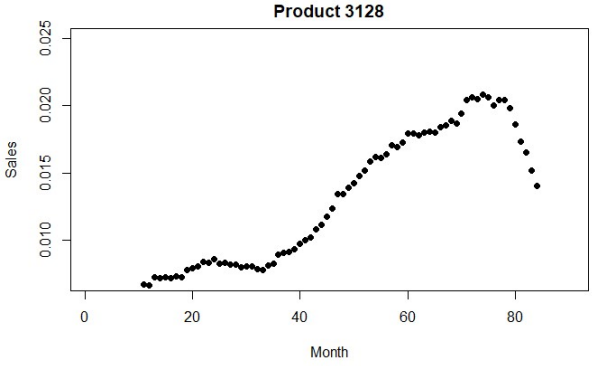
(b) US Product 3101



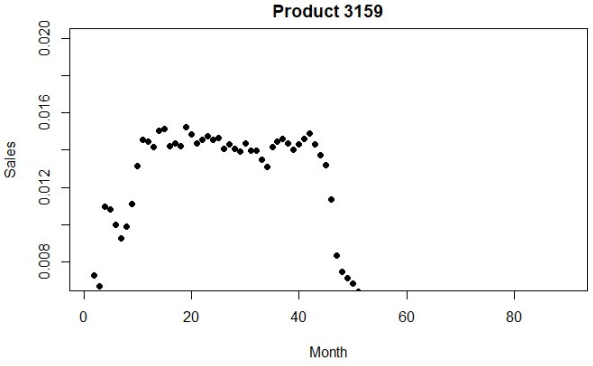
(c) US Product 3104



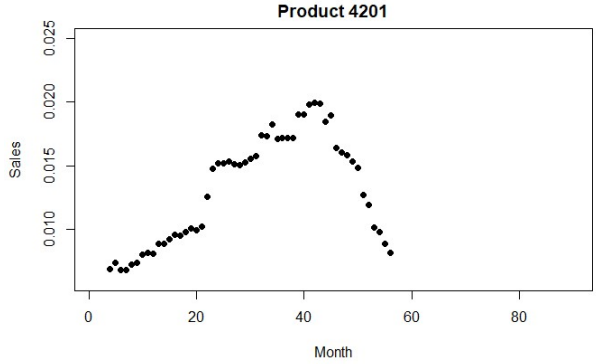
(d) US Product 3105



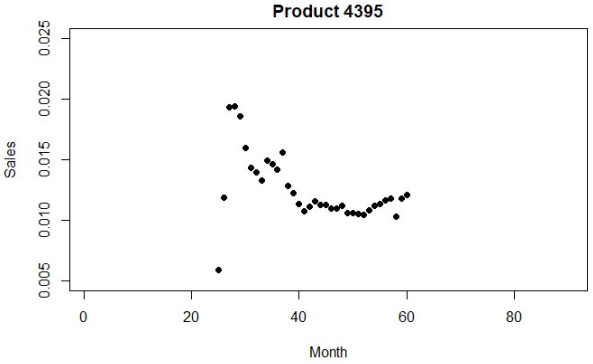
(e) US Product 3128



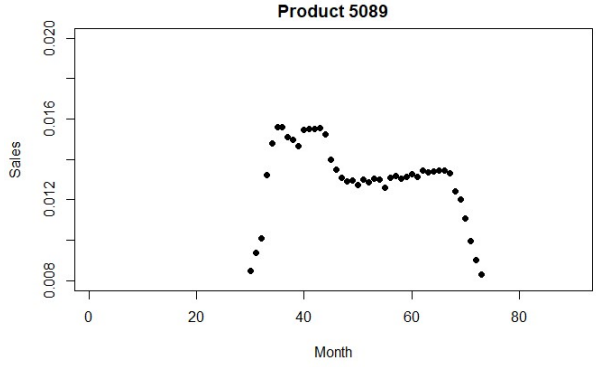
(f) US Product 3159



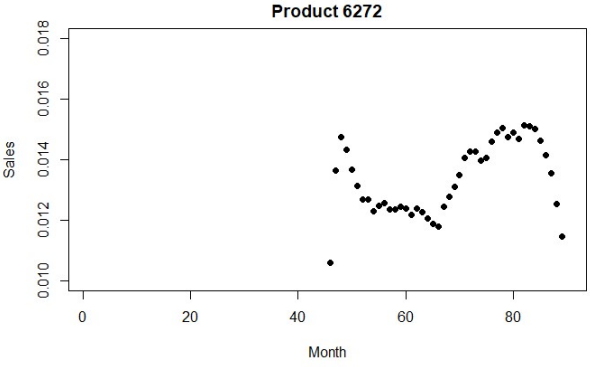
(g) US Product 4201



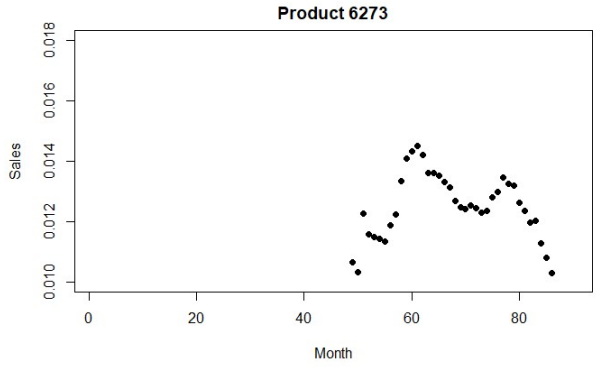
(h) US Product 4395



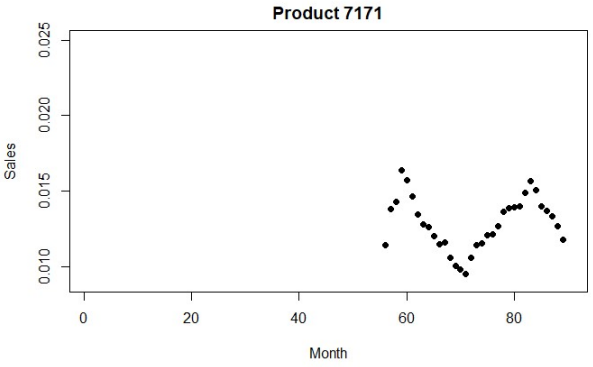
(i) US Product 5089



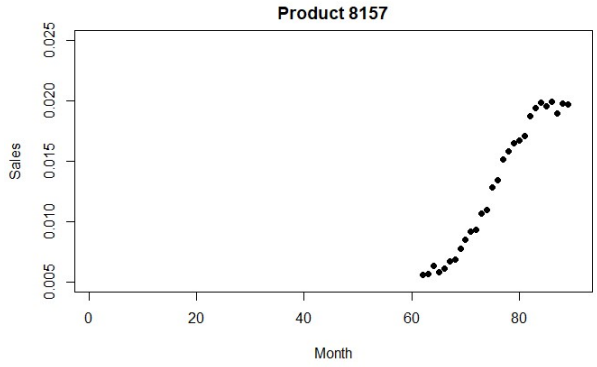
(j) US Product 6272



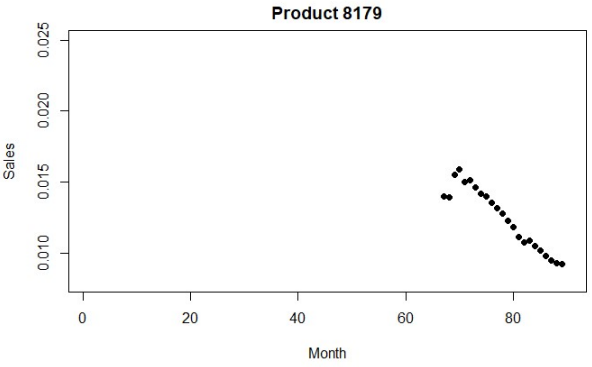
(k) US Product 6273



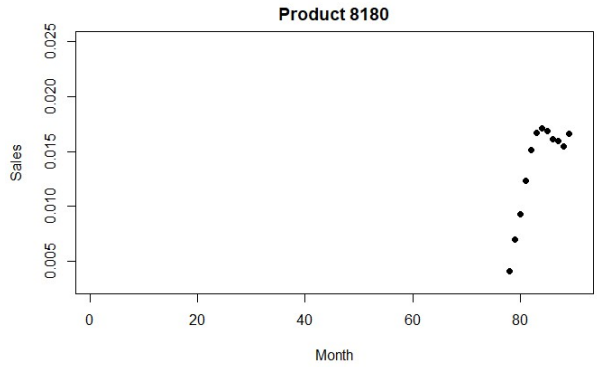
(l) US Product 7171



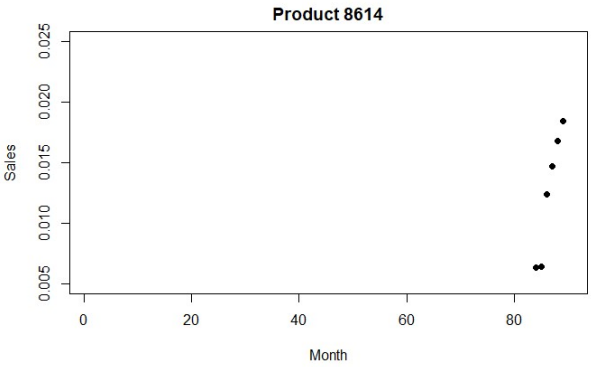
(m) US Product 8157



(n) US Product 8179



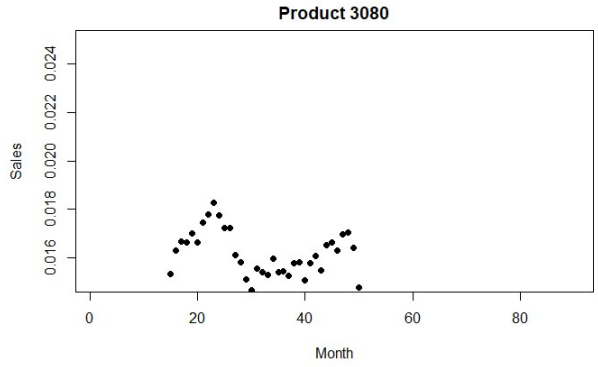
(o) US Product 8180



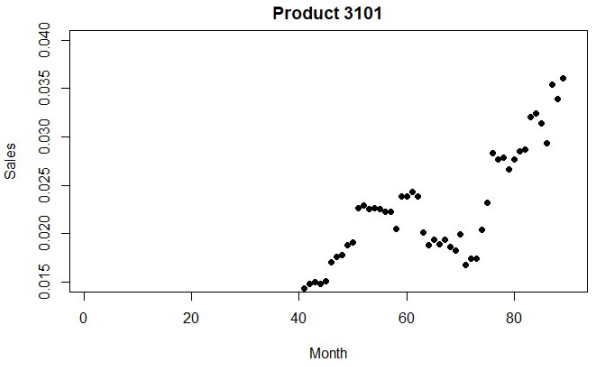
(p) US Product 8614

Figure 13: Product life cycle curves per model in the US market.

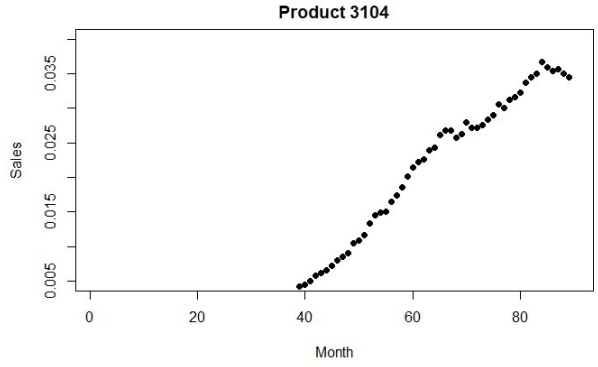
C.2 China products



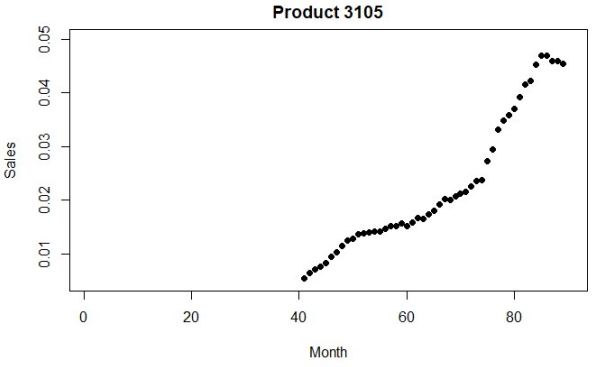
(a) China Product 3090



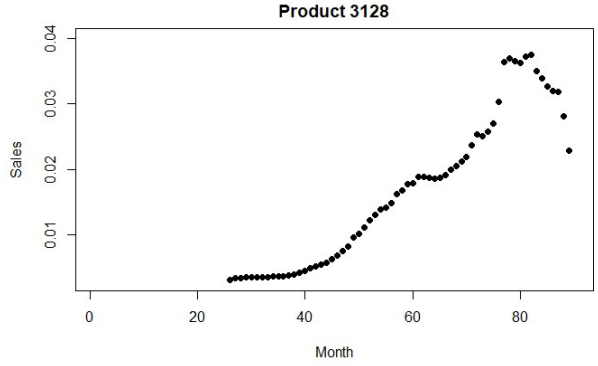
(b) China Product 3101



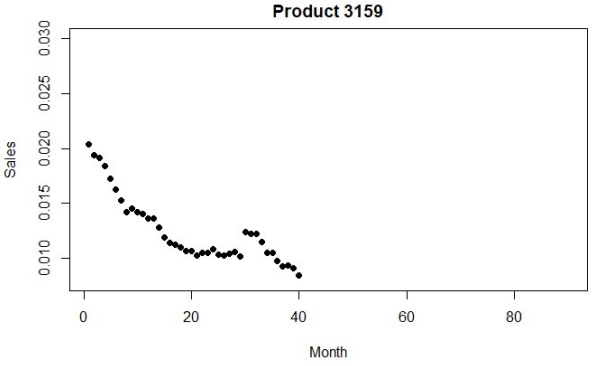
(c) China Product 3104



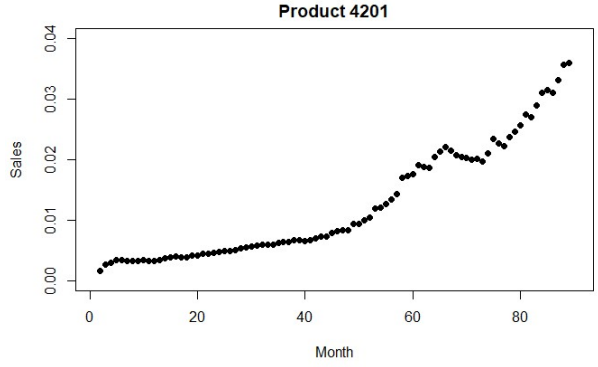
(d) China Product 3105



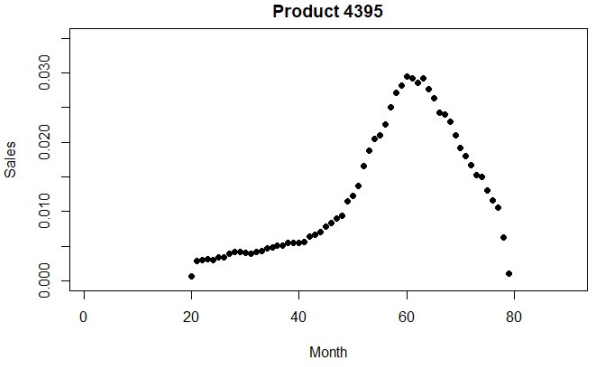
(e) China Product 3128



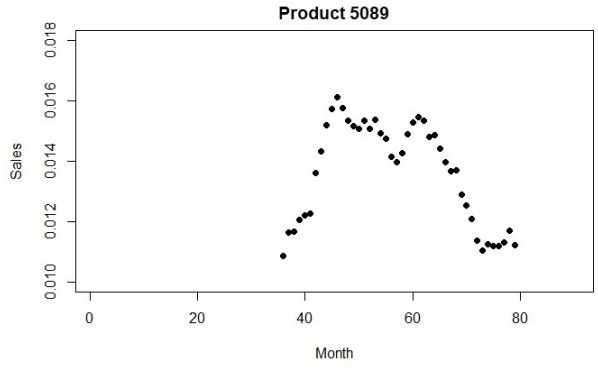
(f) China Product 3159



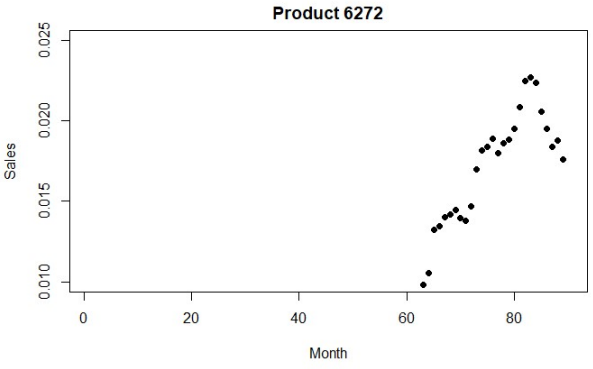
(g) China Product 4201



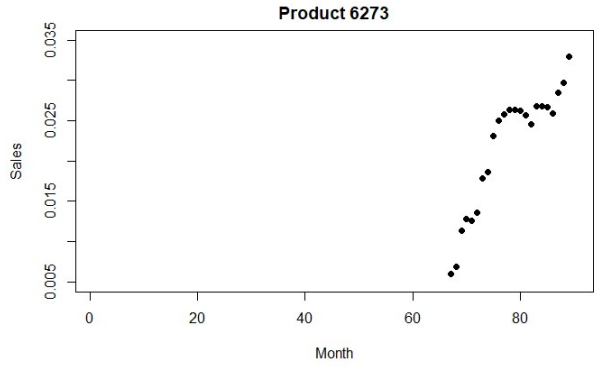
(h) China Product 4395



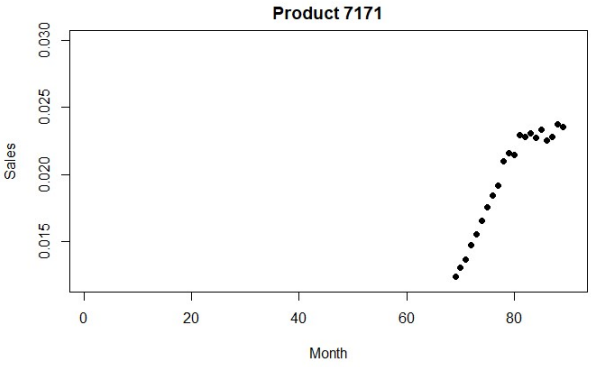
(i) China Product 5089



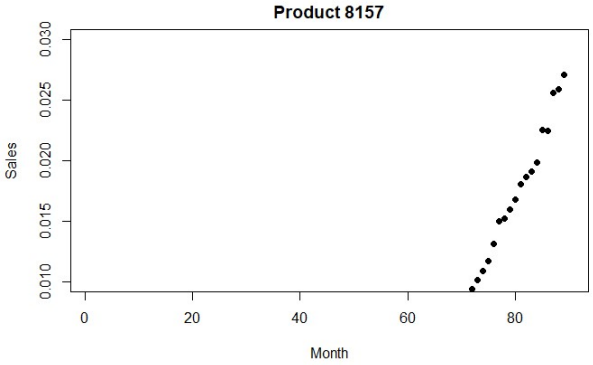
(j) China Product 6272



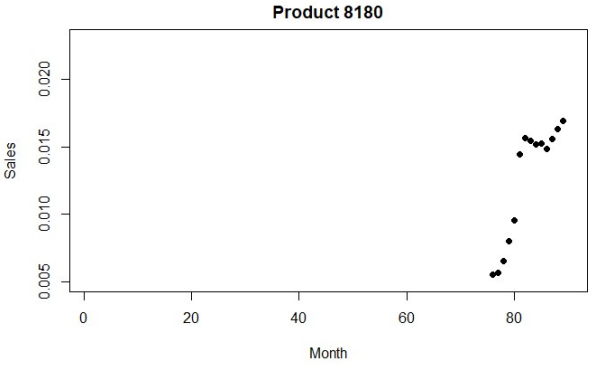
(k) China Product 6273



(l) China Product 7171



(m) China Product 8157



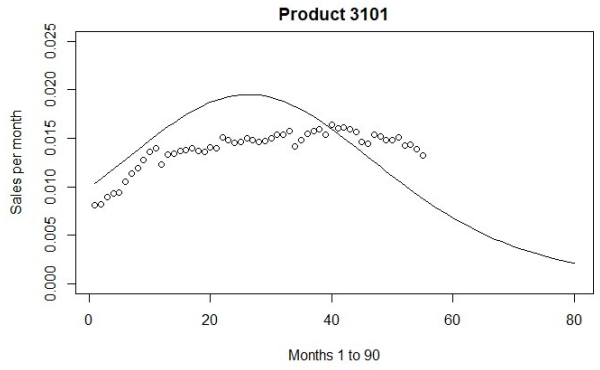
(n) China Product 8180

Figure 14: Product life cycle curves per model in the China market.

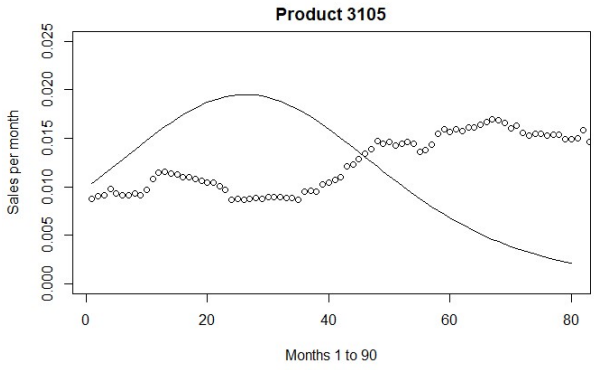
D US market models fit

This section shows graphically the how well the different models developed for the US market fit the product data.

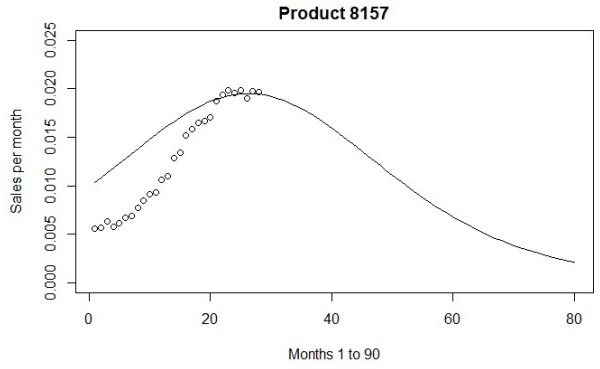
D.1 US Averaged Model (M1)



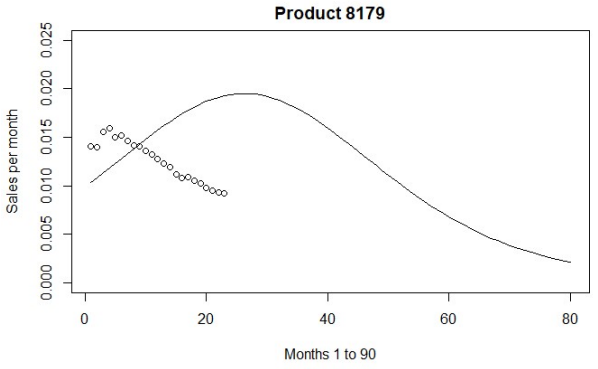
(a) M1 fit for US Product 3101



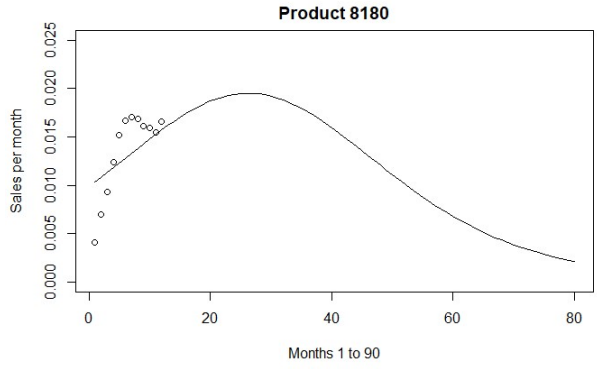
(b) M1 fit for US Product 3105



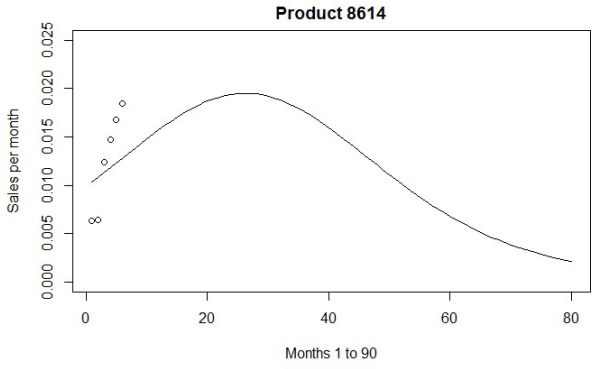
(c) M1 fit for US Product 8157



(d) M1 fit for US Product 8179



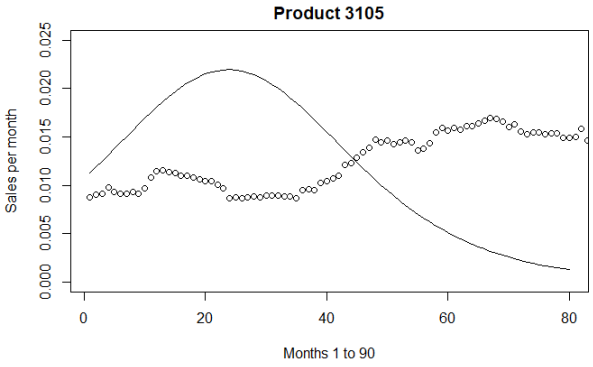
(e) M1 fit for US Product 8180



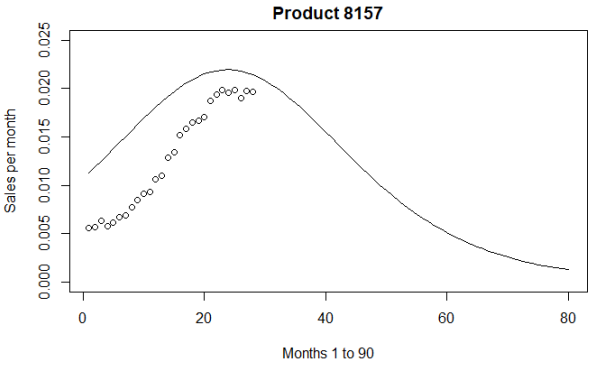
(f) M1 fit for US Product 8614

Figure 15: US Averaged Model (M1) fit in the US market.

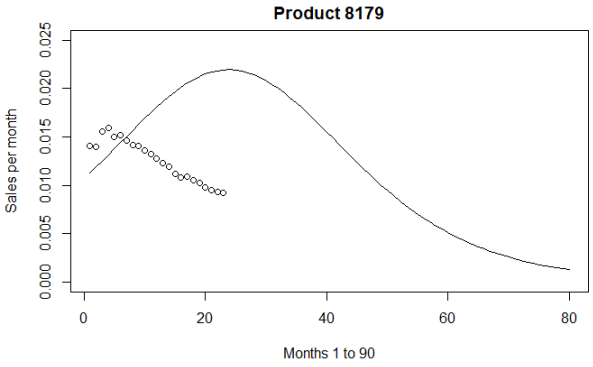
D.2 US Averaged Cluster 1 Model (M2)



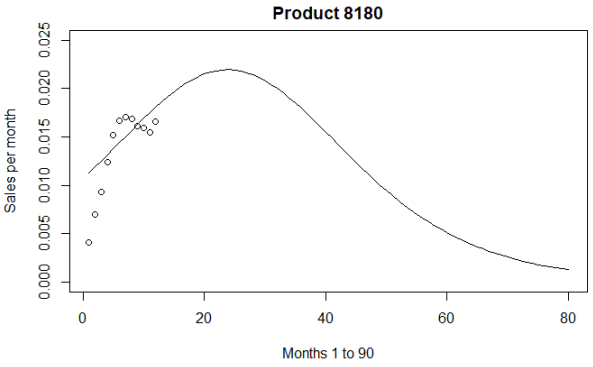
(a) M2 fit for US Product 3105



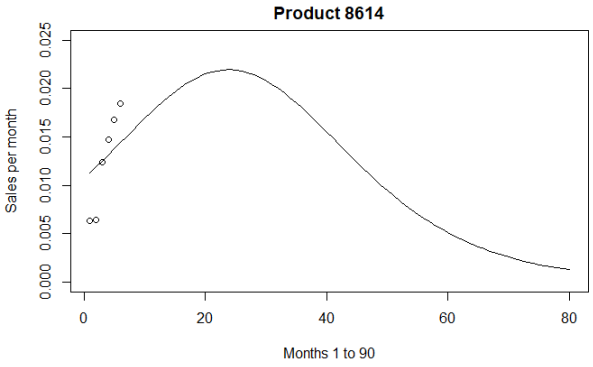
(b) M2 fit for US Product 8157



(c) M2 fit for US Product 8179



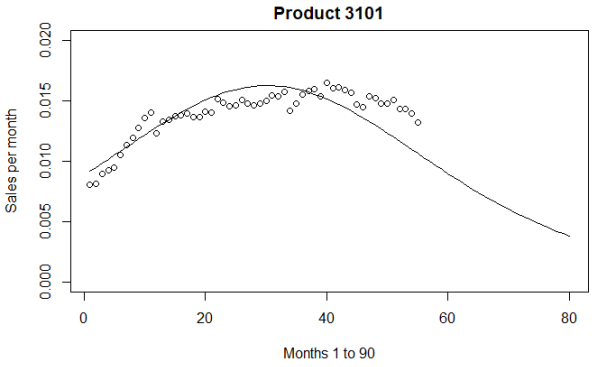
(d) M2 fit for US Product 8180



(e) M2 fit for US Product 8614

Figure 16: US Averaged Cluster 1 Model (M2) in the US market.

D.3 US Averaged Cluster 2 Model (M3)



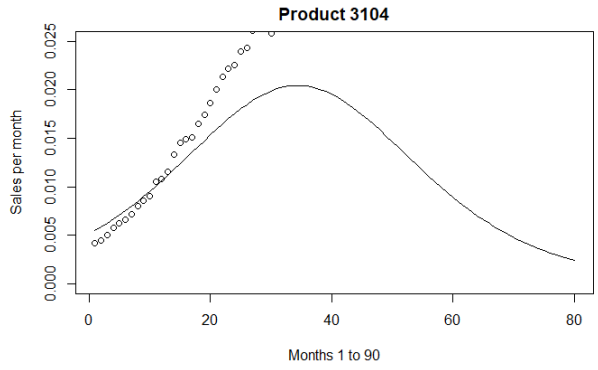
(a) M3 fit for US Product 3101

Figure 17: US Averaged Cluster 1 Model (M3) in the US market.

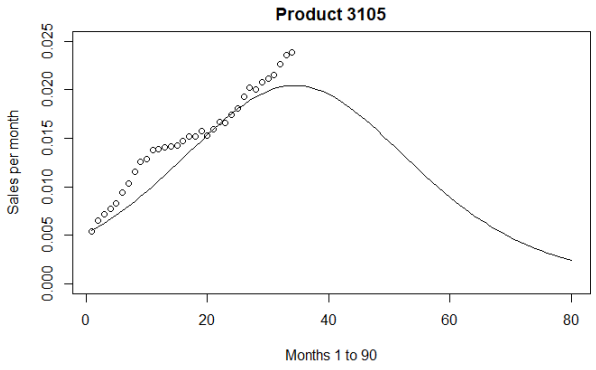
E China market models fit

This section shows graphically how well the different models developed for the China market fit the products data.

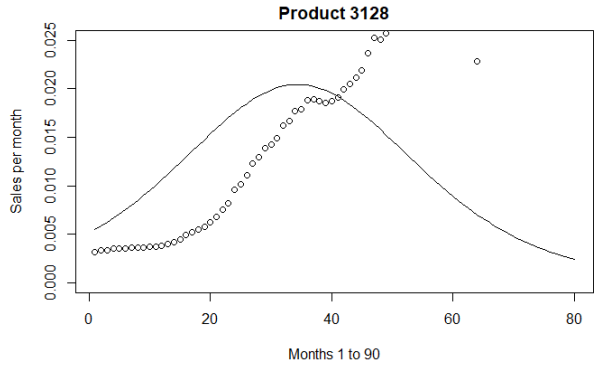
E.1 China Averaged Model (M4)



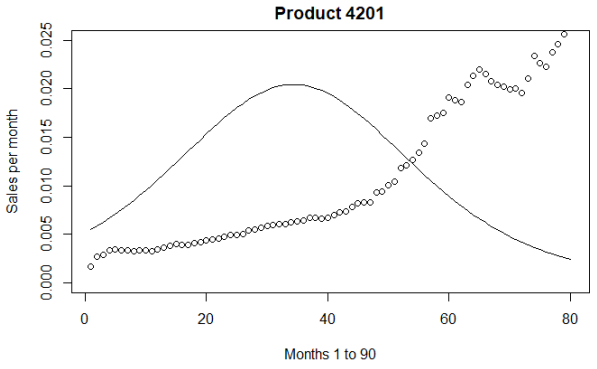
(a) M4 fit for China Product 3104



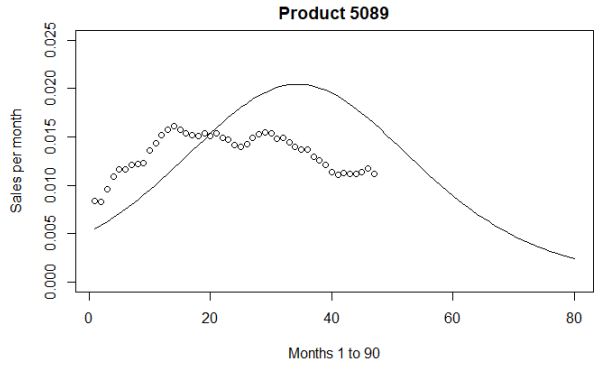
(b) M4 fit for China Product 3105



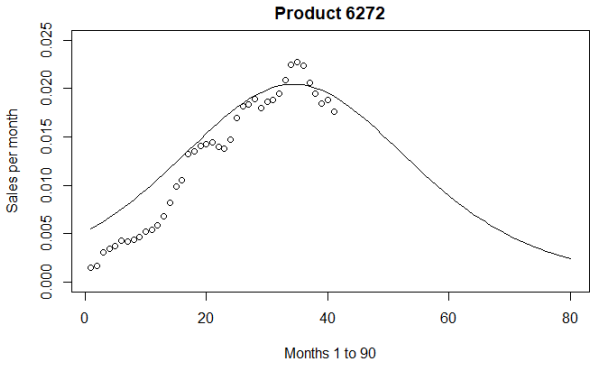
(c) M4 fit for China Product 3128



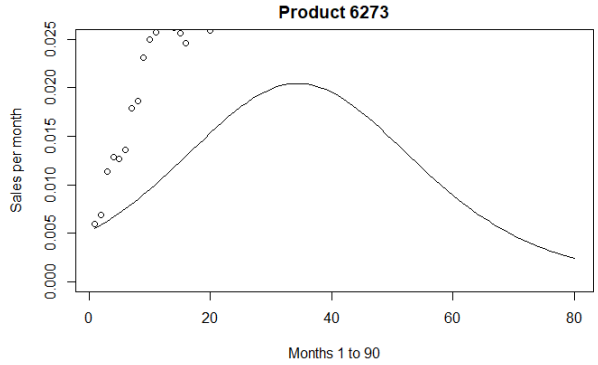
(d) M4 fit for China Product 4201



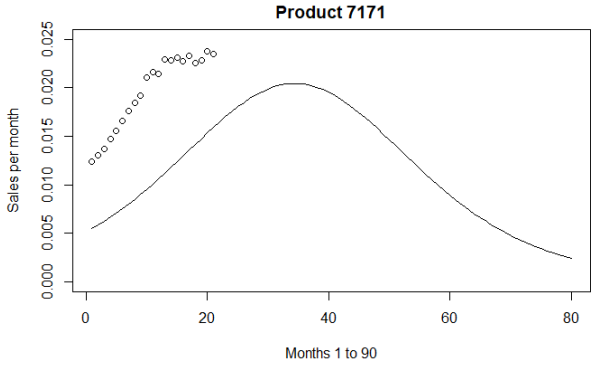
(e) M4 fit for China Product 5089



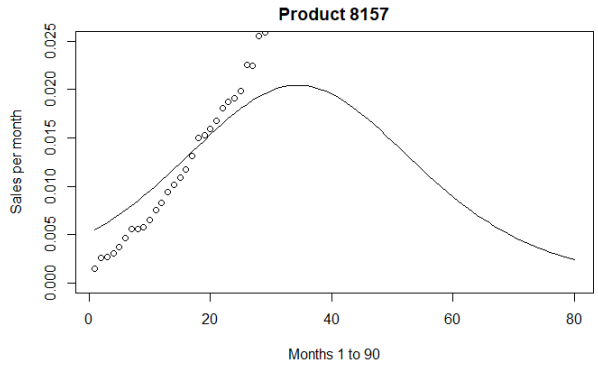
(f) M4 fit for China Product 6272



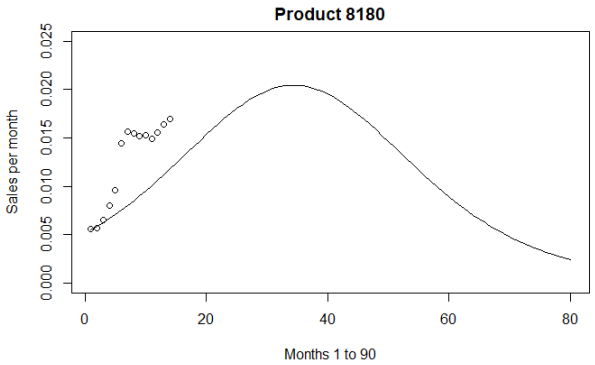
(g) M4 fit for China Product 6273



(h) M4 fit for China Product 7171



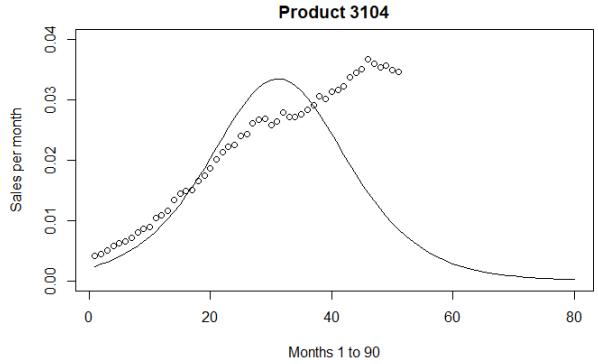
(i) M4 fit for China Product 8157



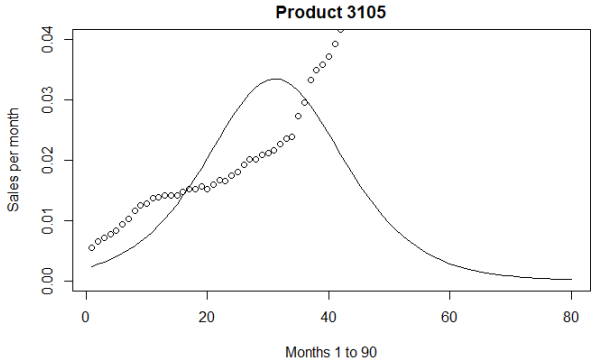
(j) M4 fit for China Product 8180

Figure 18: China Averaged Model (M4) fit in the China market.

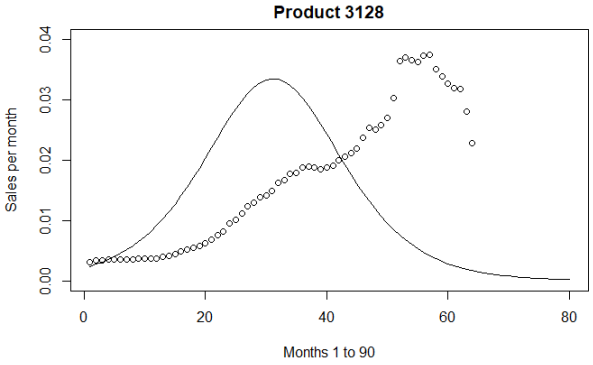
E.2 China Averaged Cluster 1 Model (M5)



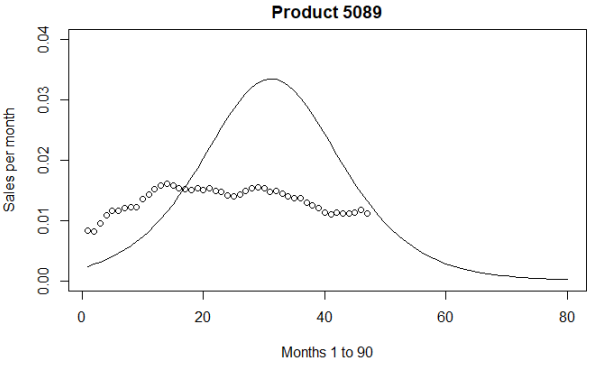
(a) M5 fit for China Product 3104



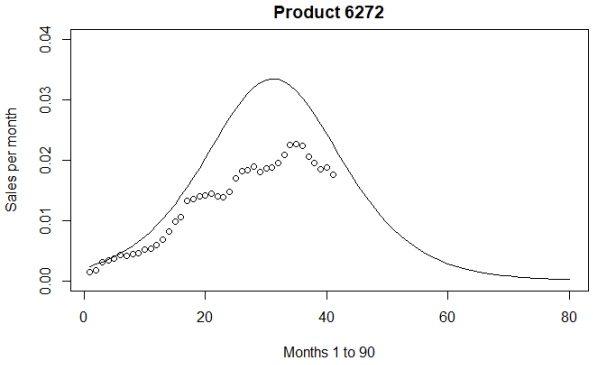
(b) M5 fit for China Product 3105



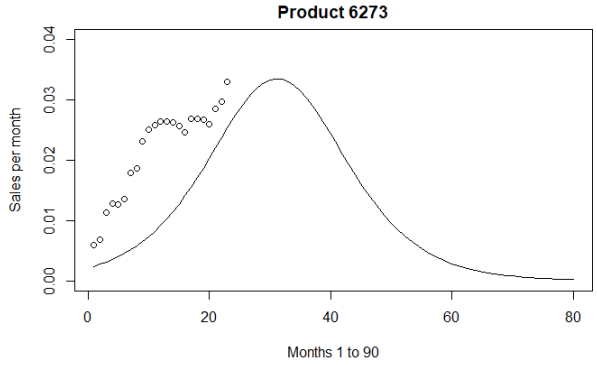
(c) M5 fit for China Product 3128



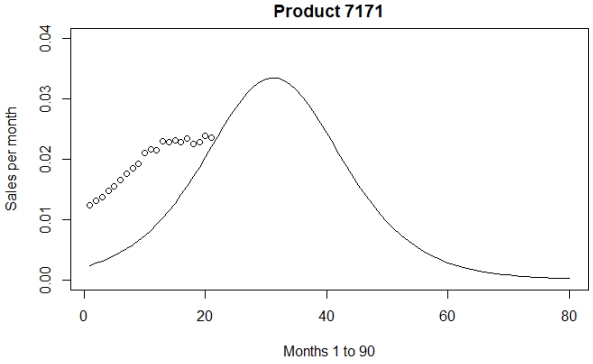
(d) M5 fit for China Product 5089



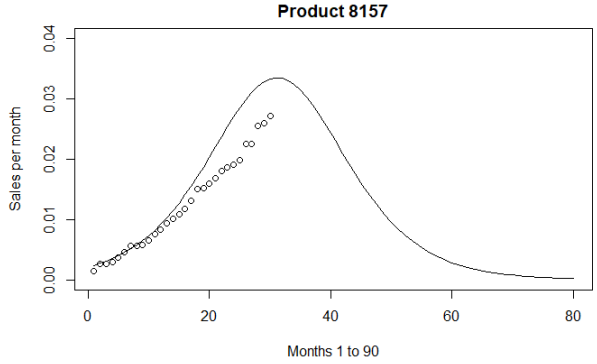
(e) M5 fit for China Product 6272



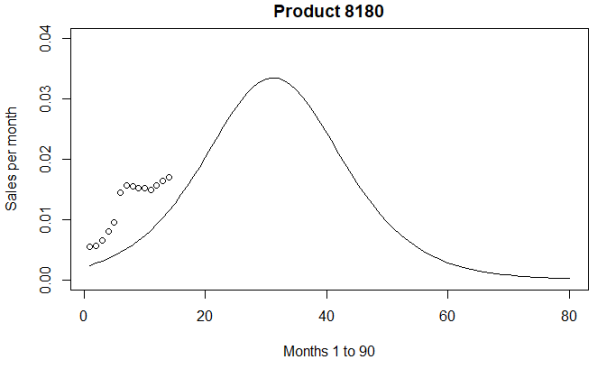
(f) M5 fit for China Product 6273



(g) M5 fit for China Product 7171



(h) M5 fit for China Product 8157



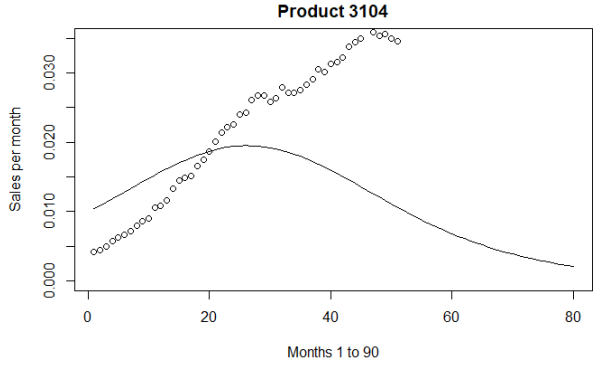
(i) M5 fit for China Product 8180

Figure 19: China Averaged Cluster 1 Model (M5) fit in the China market.

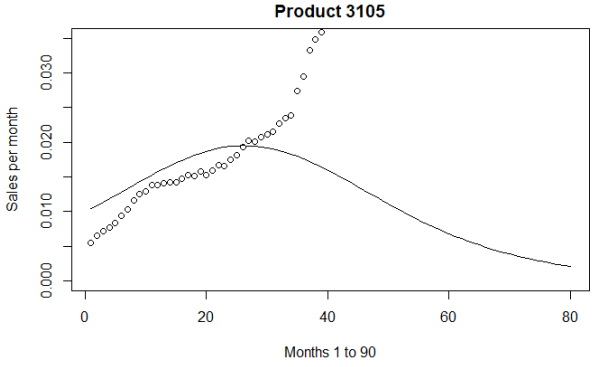
F Global market models fit

This section shows graphically the how well the different Global market models developed fit the products data.

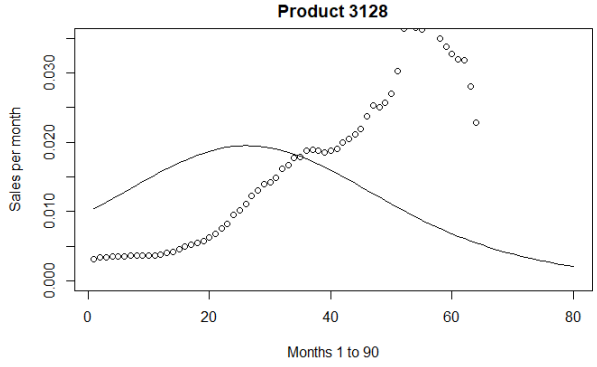
F.1 US Averaged Model (M1) in China



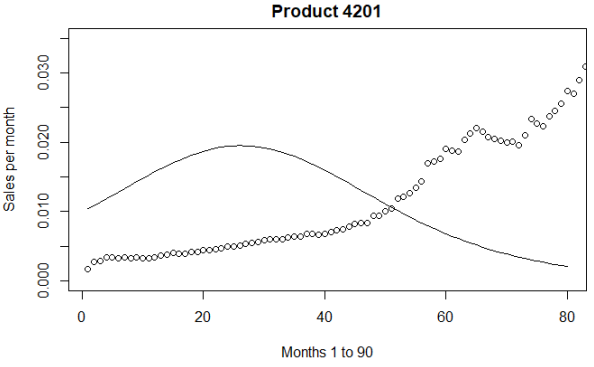
(a) M1 fit for China Product 3104



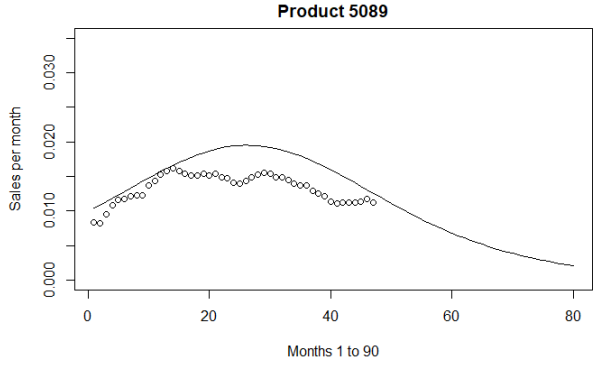
(b) M1 fit for China Product 3105



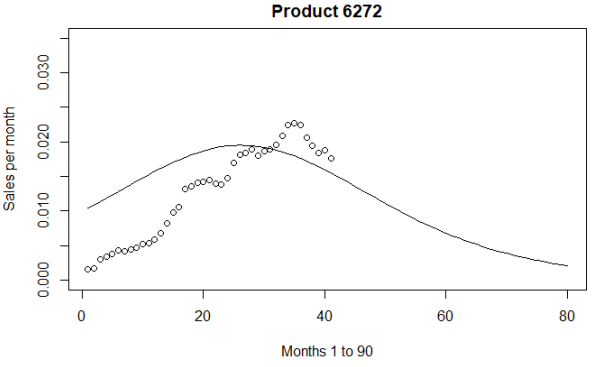
(c) M1 fit for China Product 3128



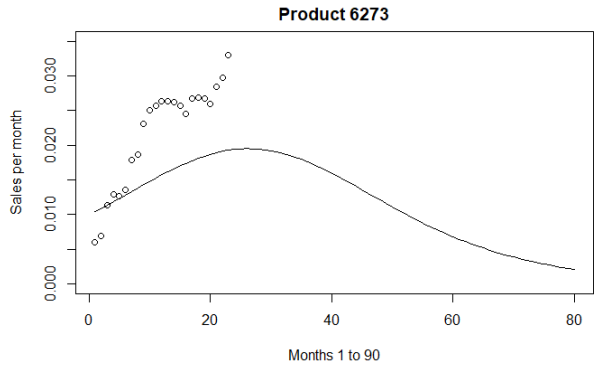
(d) M1 fit for China Product 4201



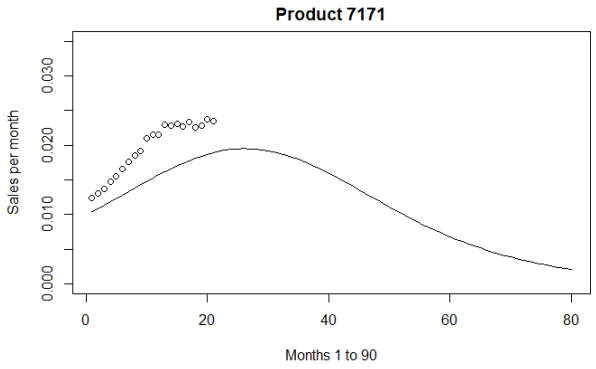
(e) M1 fit for China Product 5089



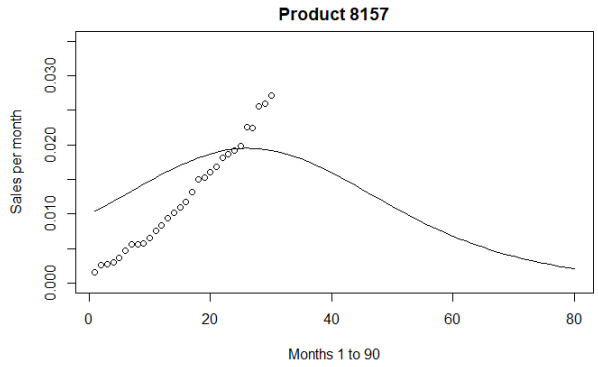
(f) M1 fit for China Product 6272



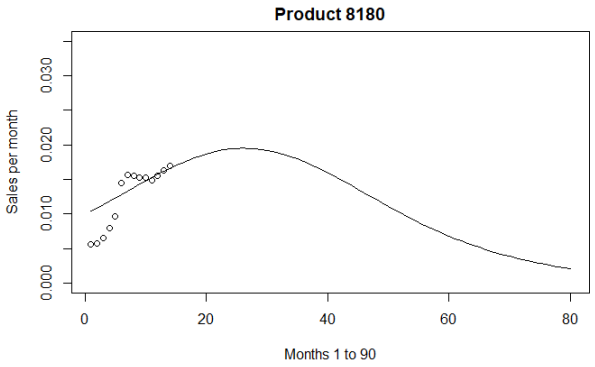
(g) M1 fit for China Product 6273



(h) M1 fit for China Product 7171



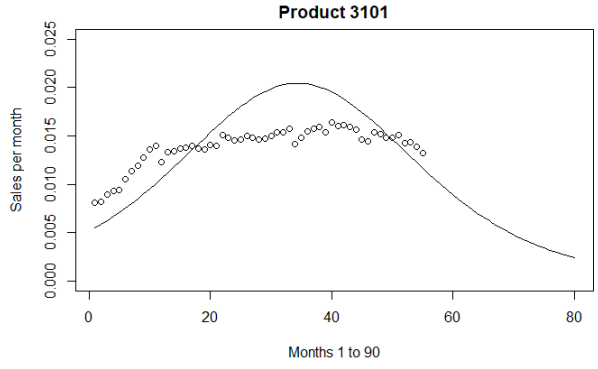
(i) M1 fit for China Product 8157



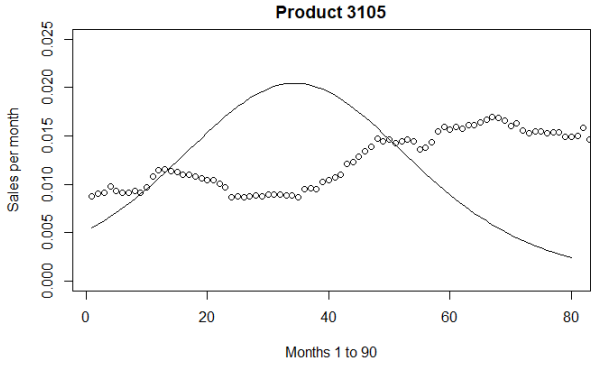
(j) M1 fit for China Product 8180

Figure 20: US Averaged Model (M1) fit in the China market.

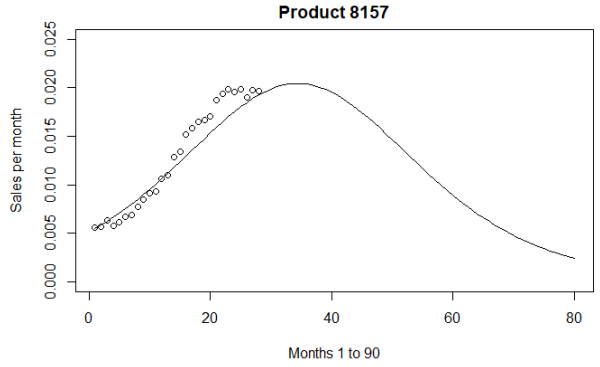
F.2 China Averaged Model (M4) in the US



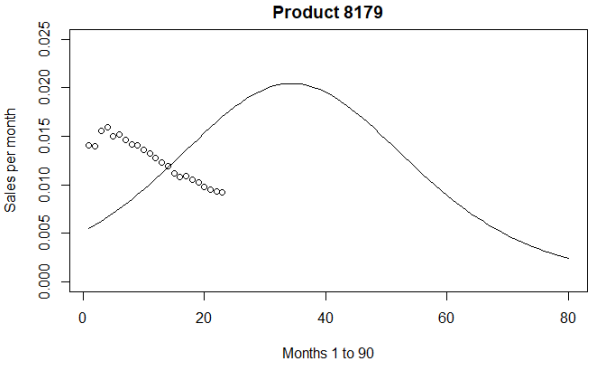
(a) M4 fit for US Product 3101



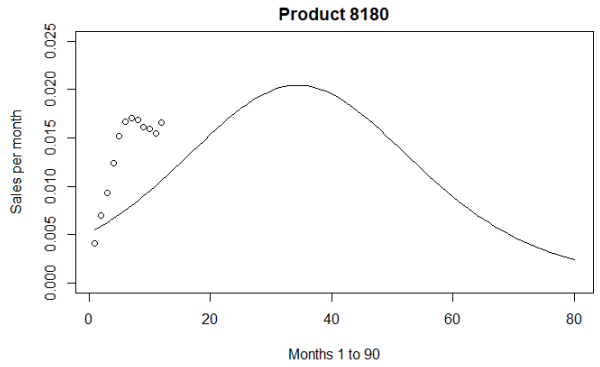
(b) M4 fit for US Product 3105



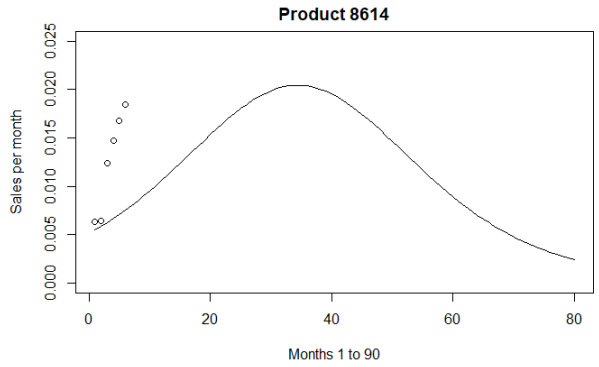
(c) M4 fit for US Product 8157



(d) M4 fit for US Product 8179



(e) M4 fit for US Product 8180

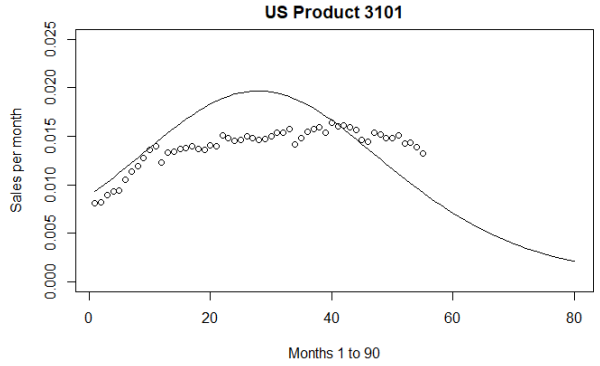


(f) M4 fit for US Product 8614

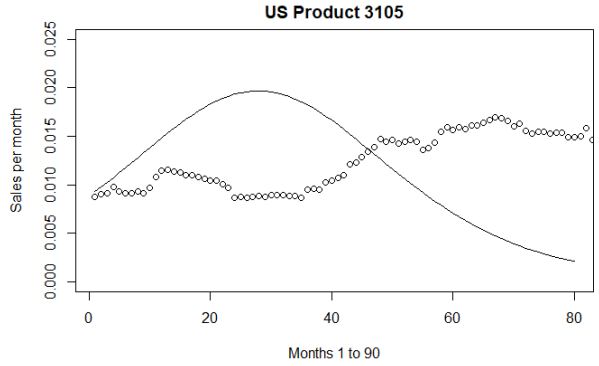
Figure 21: China Averaged Model (M4) fit in the US market.

F.3 Global Averaged Model (M7)

F.3.1 Global Averaged Model (M7) in the US

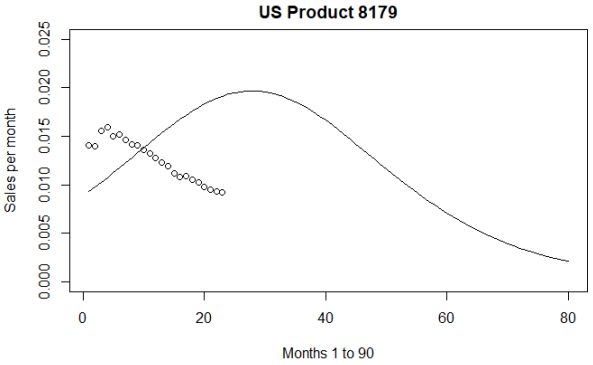


(a) M7 fit for US Product 3101

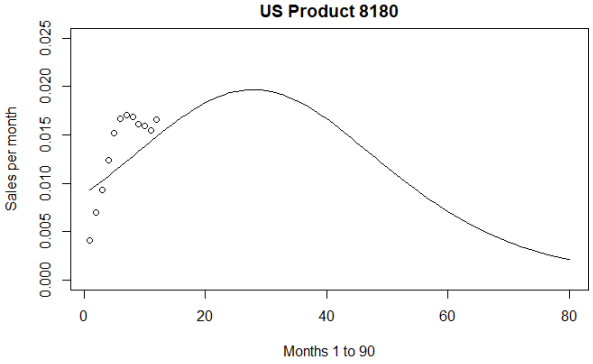


(b) M7 fit for US Product 3105

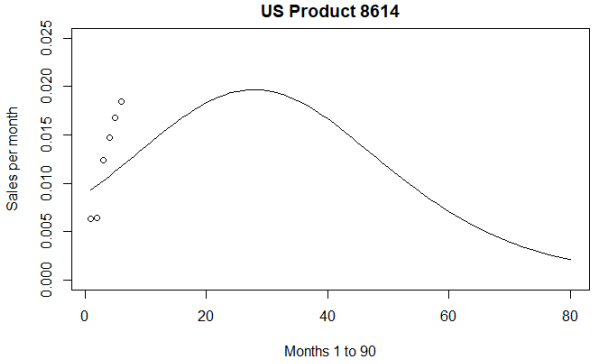
(c) M7 fit for US Product 8157



(d) M7 fit for US Product 8179



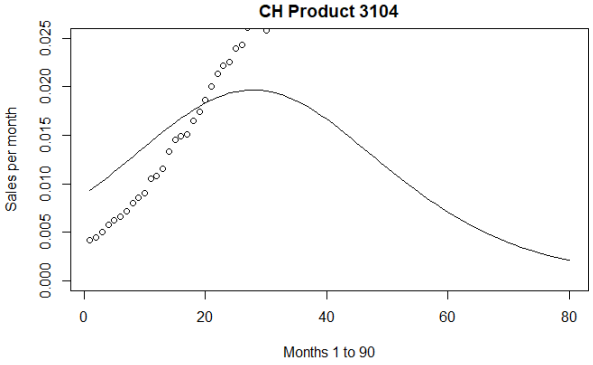
(e) M7 fit for US Product 8180



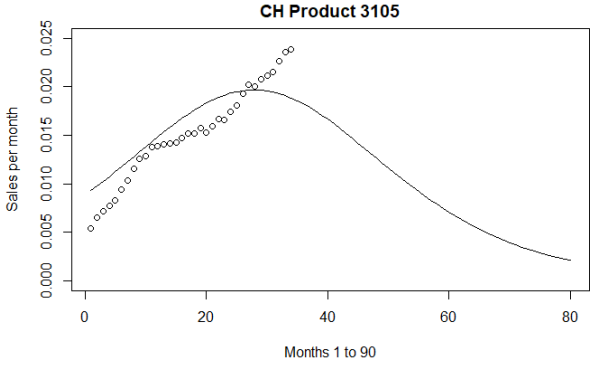
(f) M7 fit for US Product 8614

Figure 22: Global Averaged Model (M7) fit in the US market.

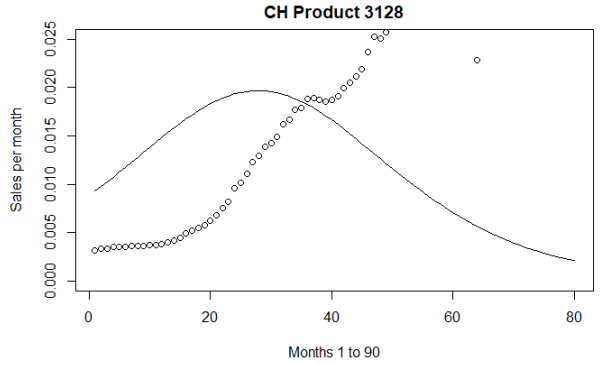
F.3.2 Global Averaged Model (M7) in China



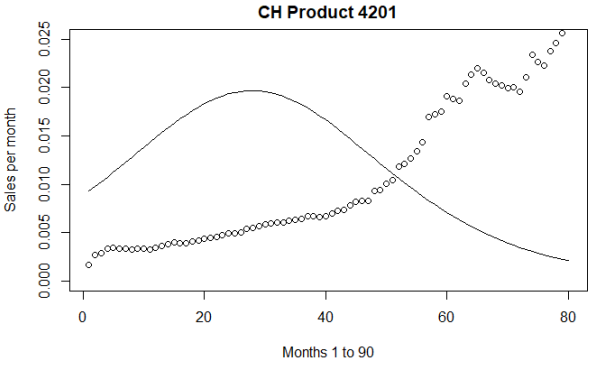
(a) M7 fit for China Product 3104



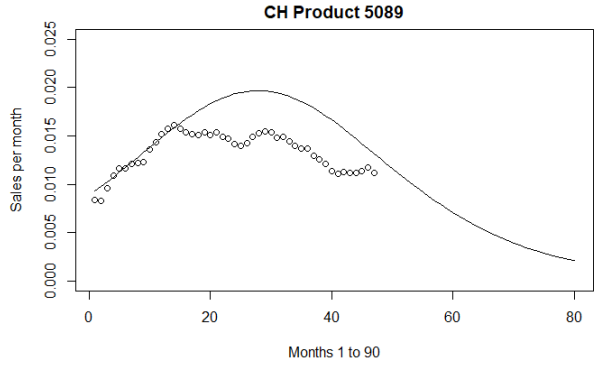
(b) M7 fit for China Product 3105



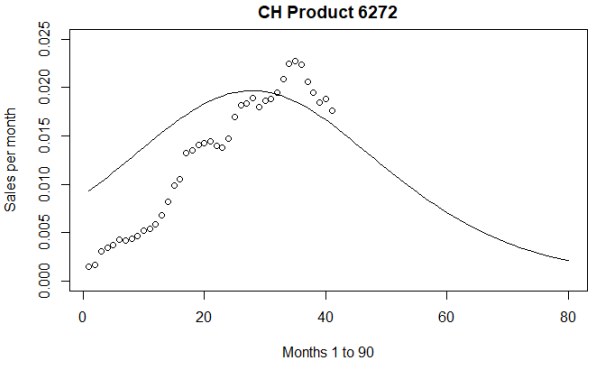
(c) M7 fit for China Product 3128



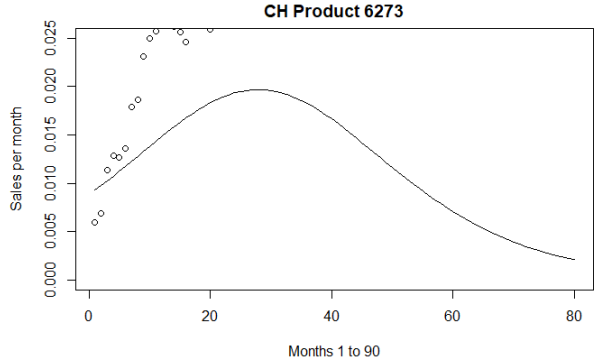
(d) M7 fit for China Product 4201



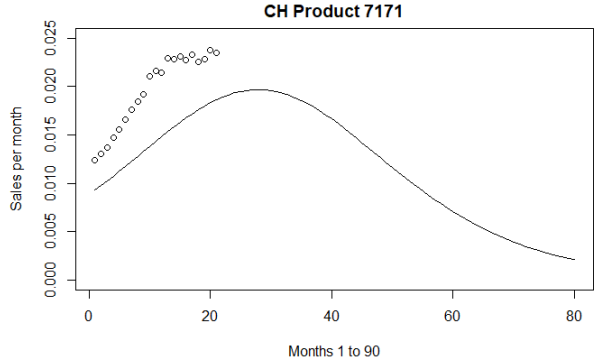
(e) M7 fit for China Product 5089



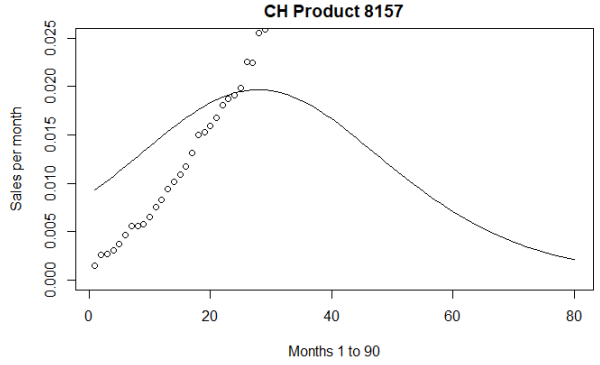
(f) M7 fit for China Product 6272



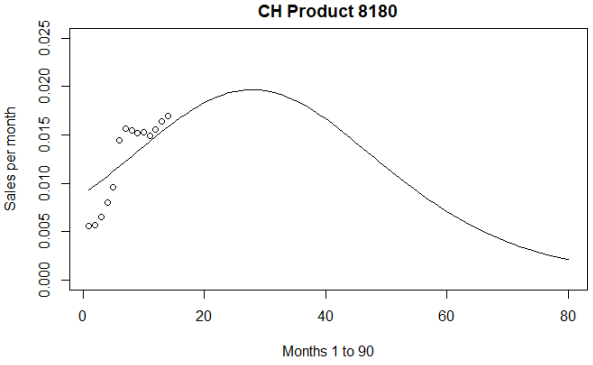
(g) M7 fit for China Product 6273



(h) M7 fit for China Product 7171



(i) M7 fit for China Product 8157



(j) M7 fit for China Product 8180

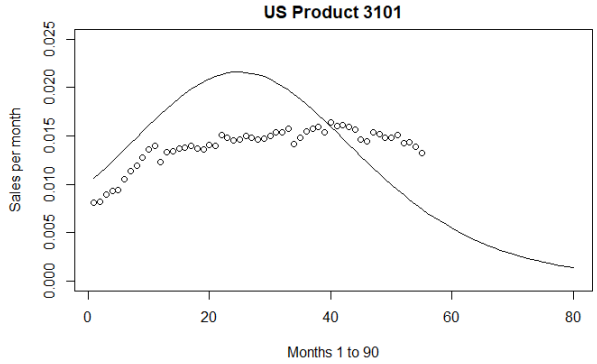
Figure 23: Global Averaged Model (M7) fit in the China market.

G Total Sales models fit

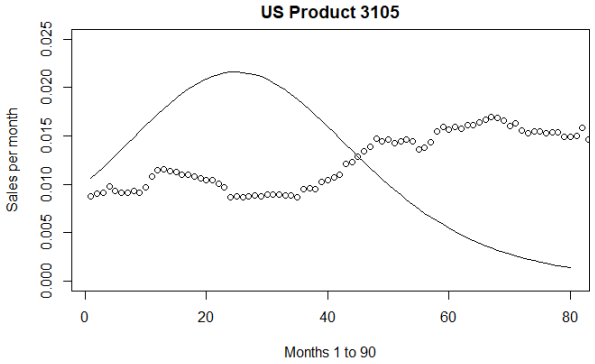
This section shows graphically the how well the different Total Sales models developed fit the products data.

G.1 US Total Sales Model (M8)

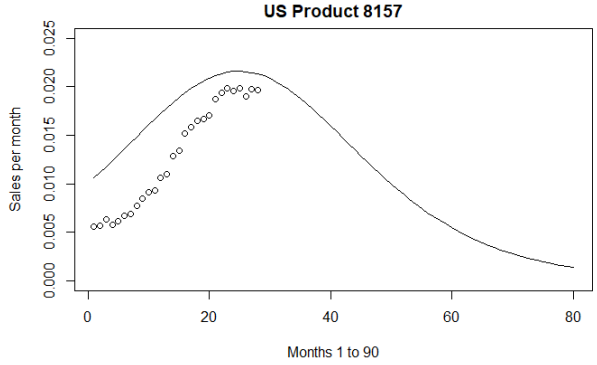
G.1.1 US Total Sales Model (M8) fit in the US



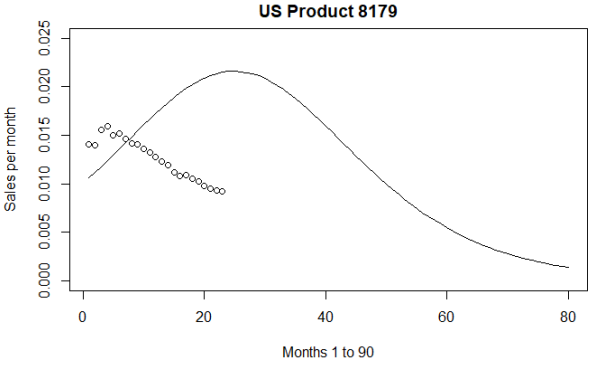
(a) M8 fit for US Product 3101



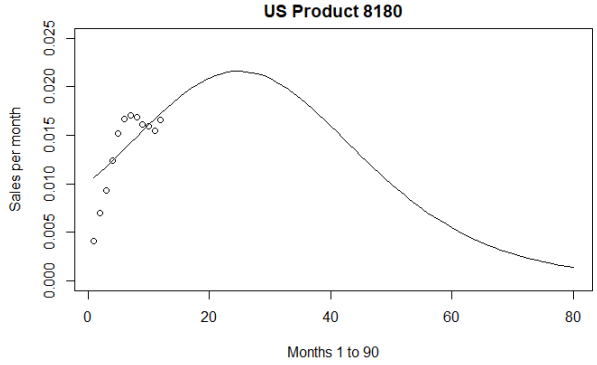
(b) M8 fit for US Product 3105



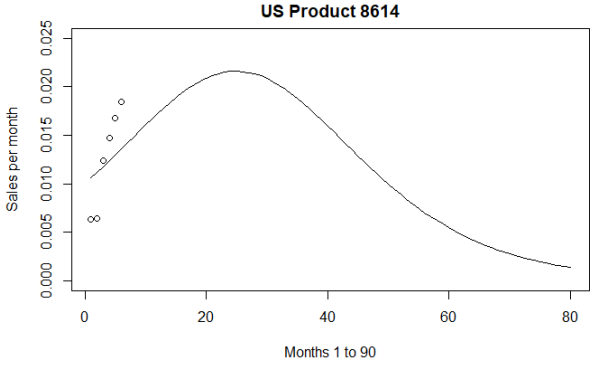
(c) M8 fit for US Product 8157



(d) M8 fit for US Product 8179



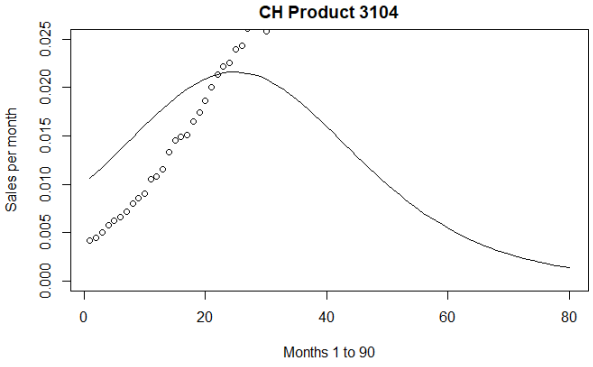
(e) M8 fit for US Product 8180



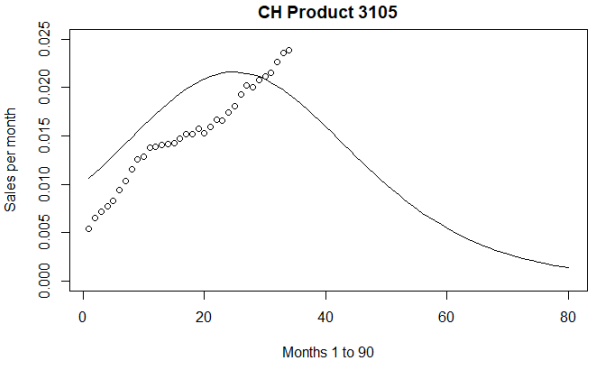
(f) M8 fit for US Product 8614

Figure 24: US Total Sales Model (M8) fit in the US market.

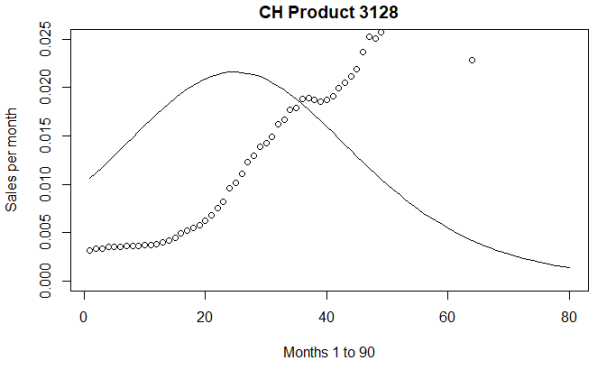
G.1.2 US Total Sales Model (M8) fit in China



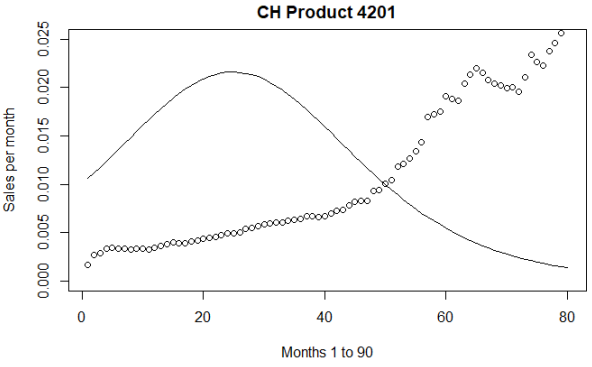
(a) M8 fit for China Product 3104



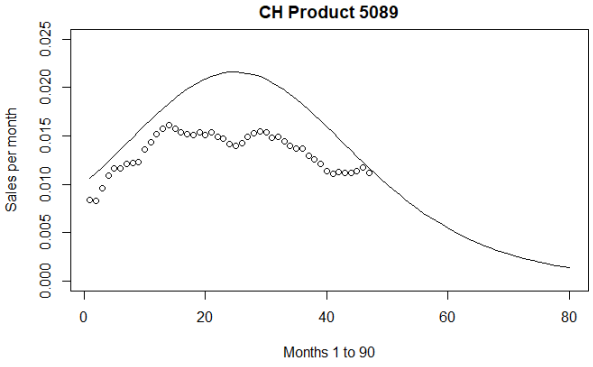
(b) M8 fit for China Product 3105



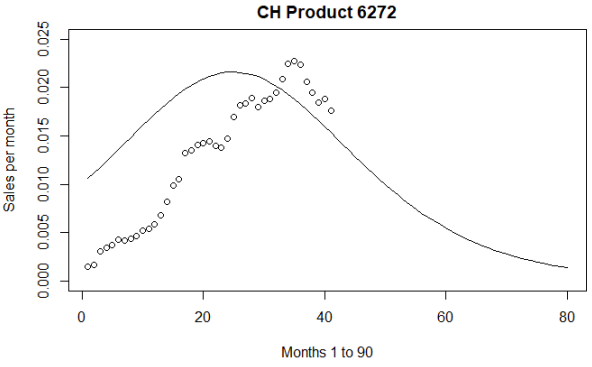
(c) M8 fit for China Product 3128



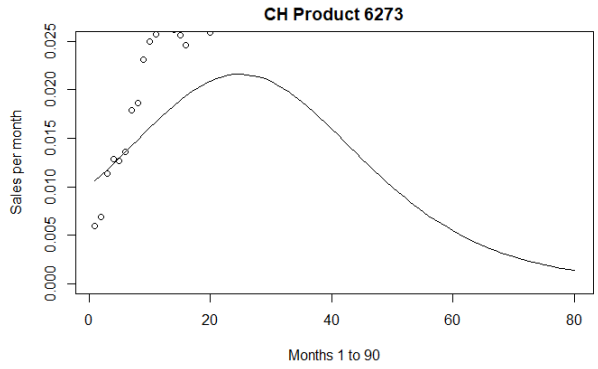
(d) M8 fit for China Product 4201



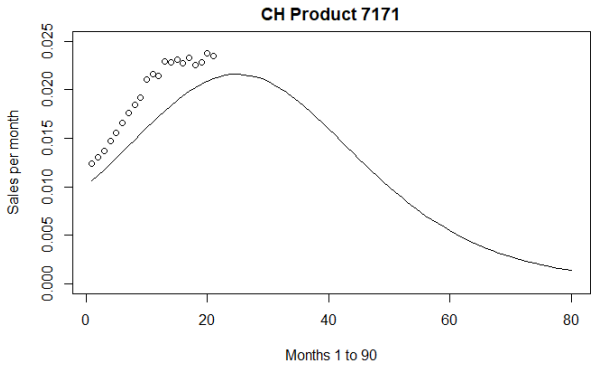
(e) M8 fit for China Product 5089



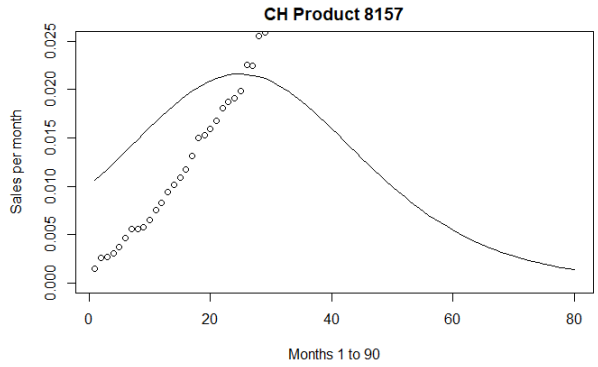
(f) M8 fit for China Product 6272



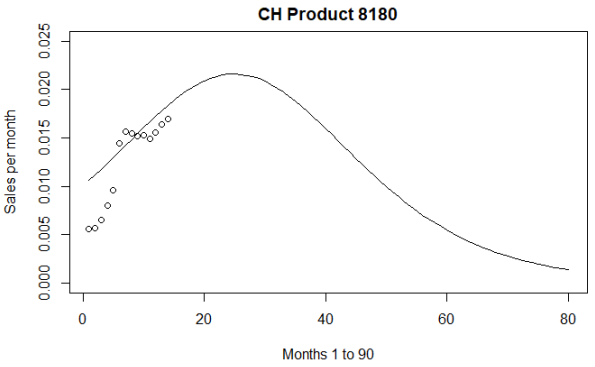
(g) M8 fit for China Product 6273



(h) M8 fit for China Product 7171



(i) M8 fit for China Product 8157

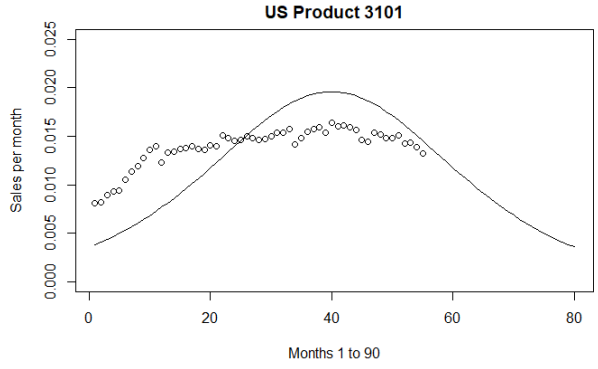


(j) M8 fit for China Product 8180

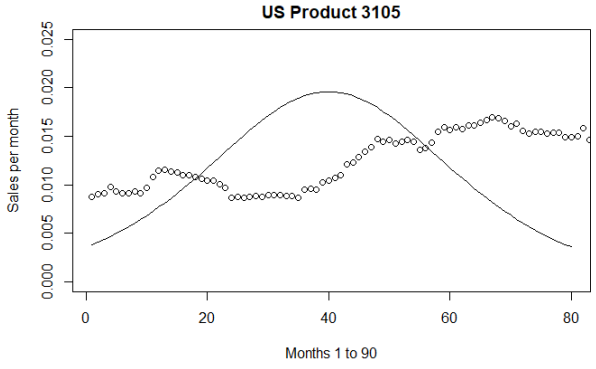
Figure 25: US Total Sales Model (M8) fit in the China market.

G.2 China Total Sales Model (M9)

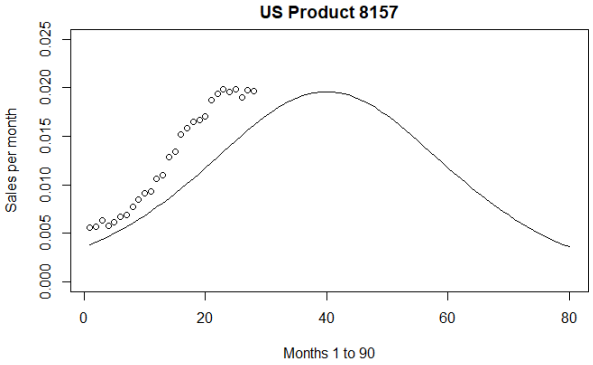
G.2.1 China Total Sales Model (M9) fit in the US



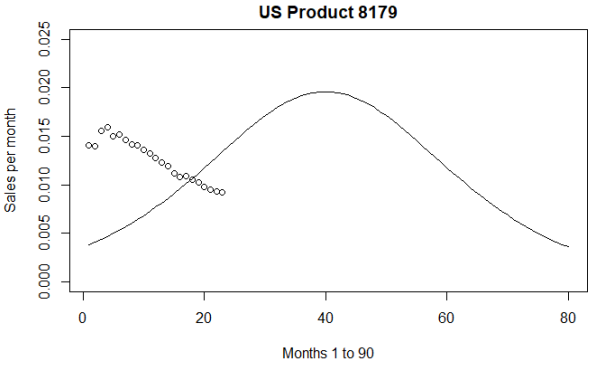
(a) M9 fit for US Product 3101



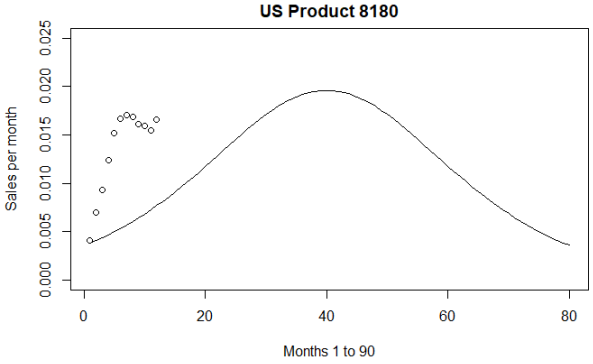
(b) M9 fit for US Product 3105



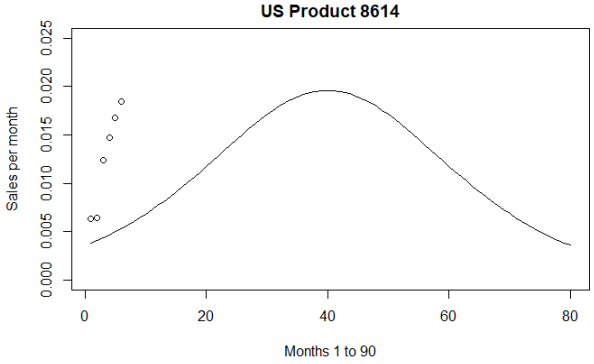
(c) M9 fit for US Product 8157



(d) M9 fit for US Product 8179



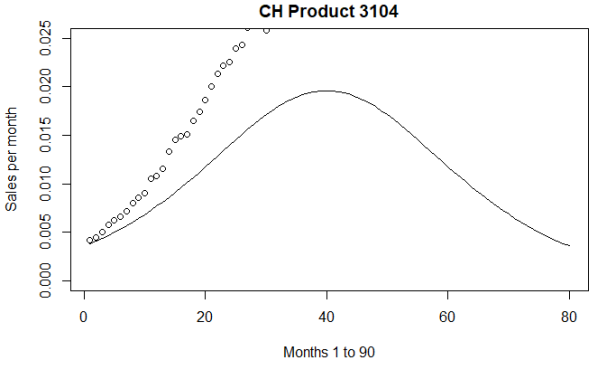
(e) M9 fit for US Product 8180



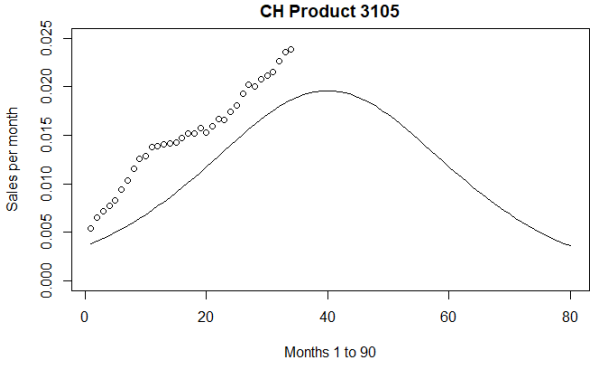
(f) M9 fit for US Product 8614

Figure 26: China Total Sales Model (M9) fit in the US market.

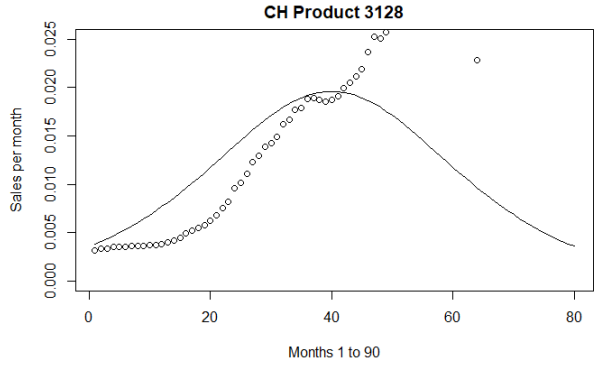
G.2.2 China Total Sales Model (M9) fit in China



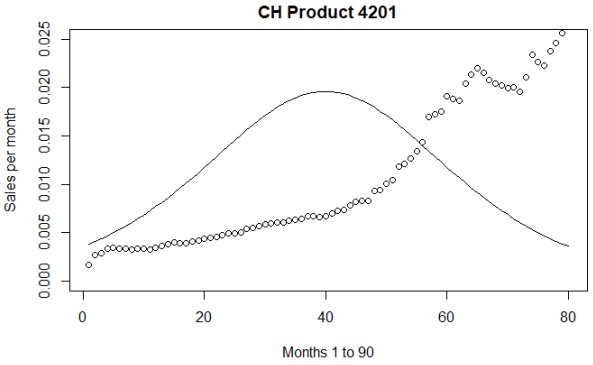
(a) M9 fit for China Product 3104



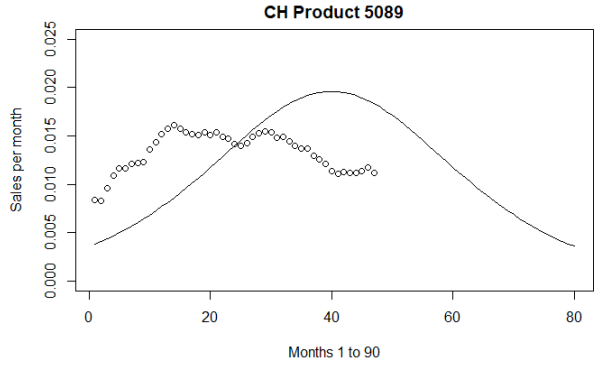
(b) M9 fit for China Product 3105



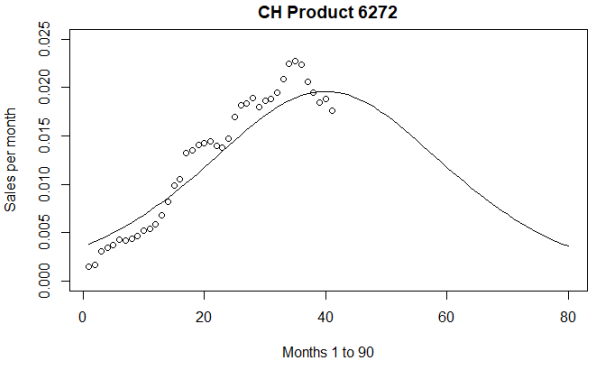
(c) M9 fit for China Product 3128



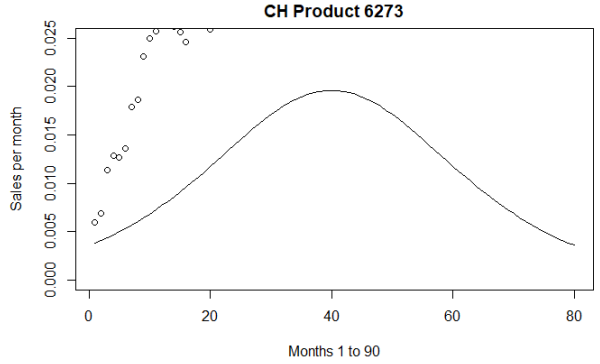
(d) M9 fit for China Product 4201



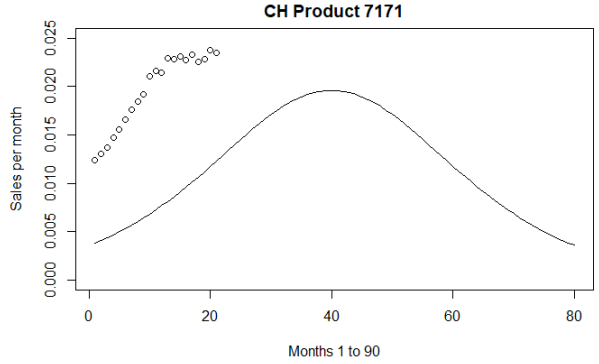
(e) M9 fit for China Product 5089



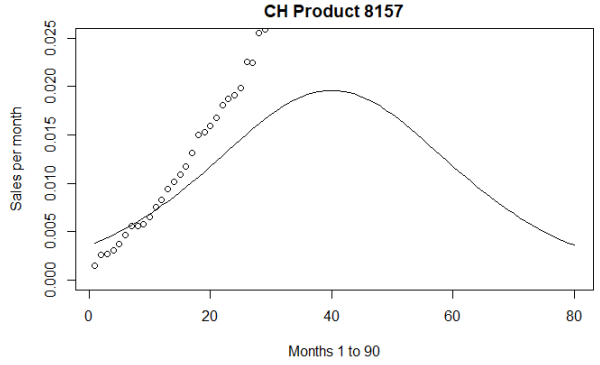
(f) M9 fit for China Product 6272



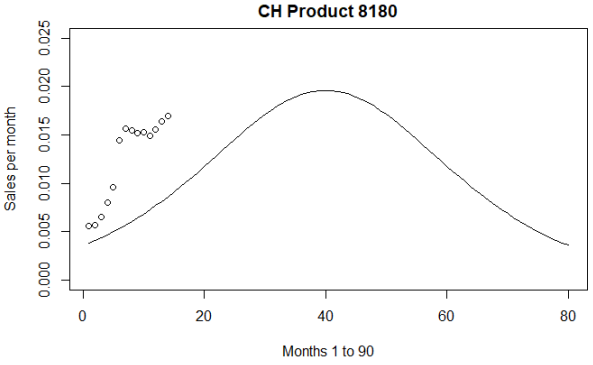
(g) M9 fit for China Product 6273



(h) M9 fit for China Product 7171



(i) M9 fit for China Product 8157

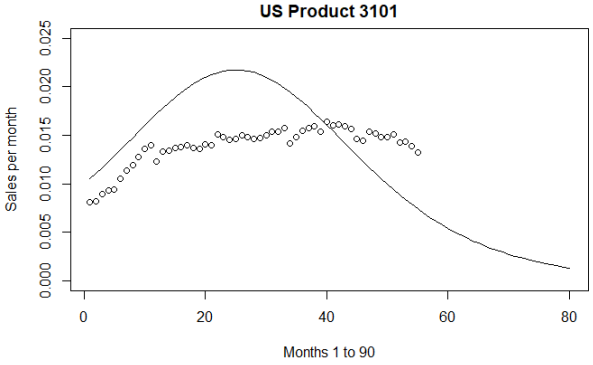


(j) M9 fit for China Product 8180

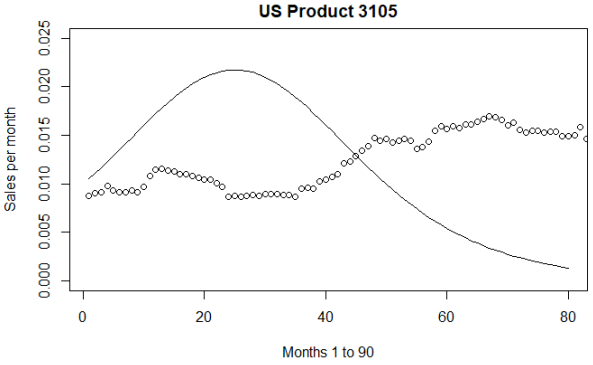
Figure 27: China Total Sales Model (M9) fit in the China market.

G.3 Global Total Sales Model (M10)

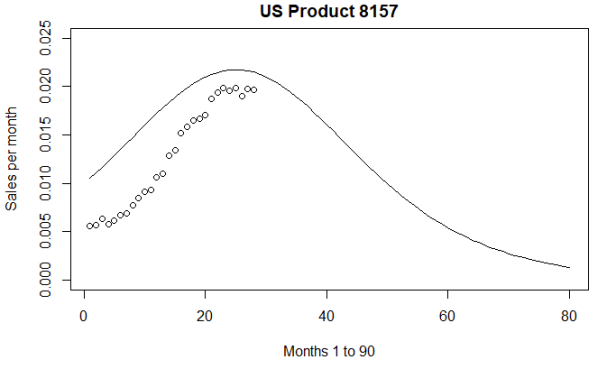
G.3.1 Global Total Sales Model (M10) fit in the US



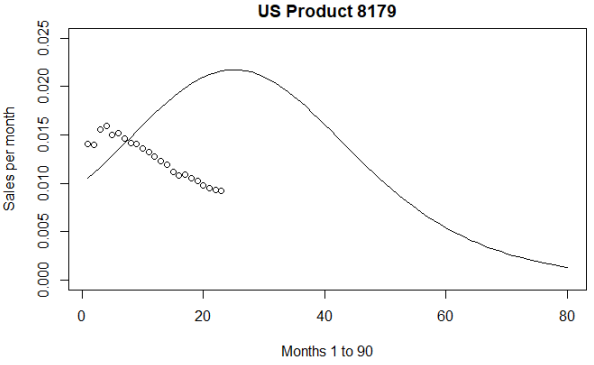
(a) M10 fit for US Product 3101



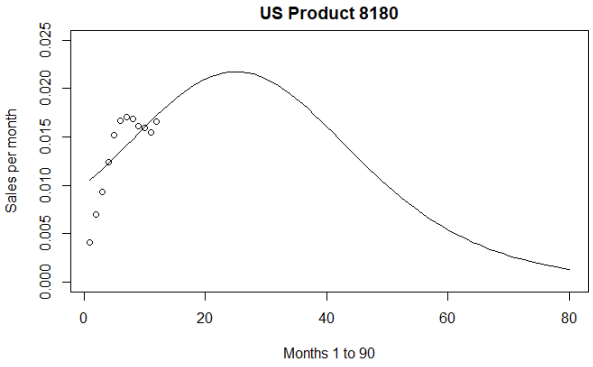
(b) M10 fit for US Product 3105



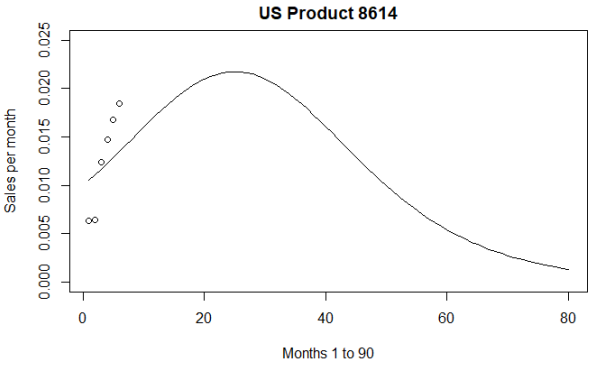
(c) M10 fit for US Product 8157



(d) M10 fit for US Product 8179



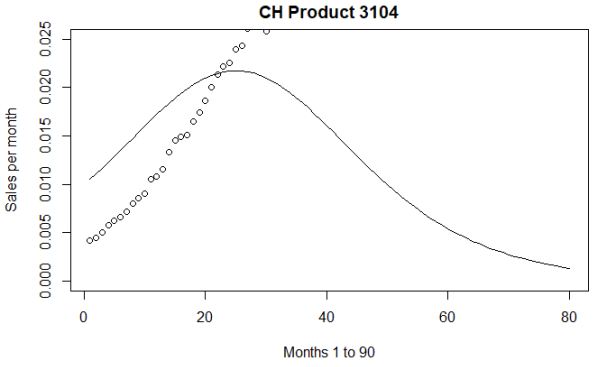
(e) M10 fit for US Product 8180



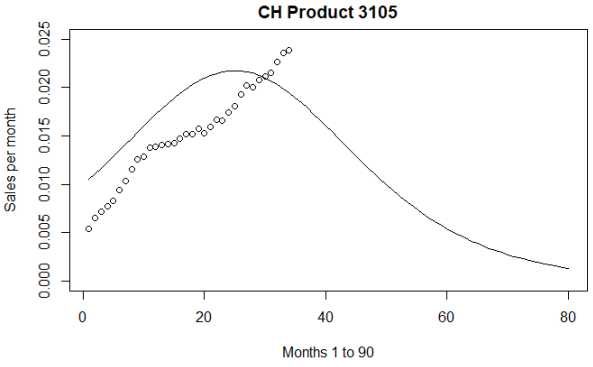
(f) M10 fit for US Product 8614

Figure 28: Global Total Sales Model (M10) fit in the US market.

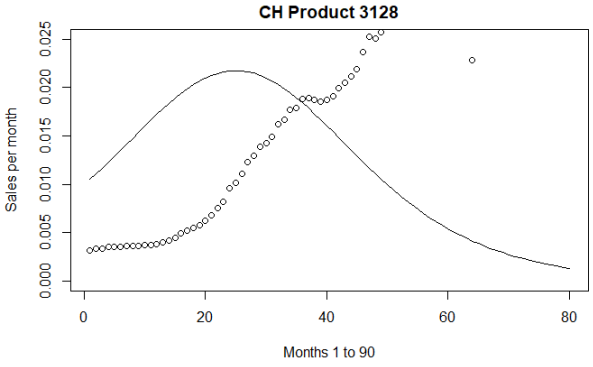
G.3.2 Global Total Sales Model (M10) fit in China



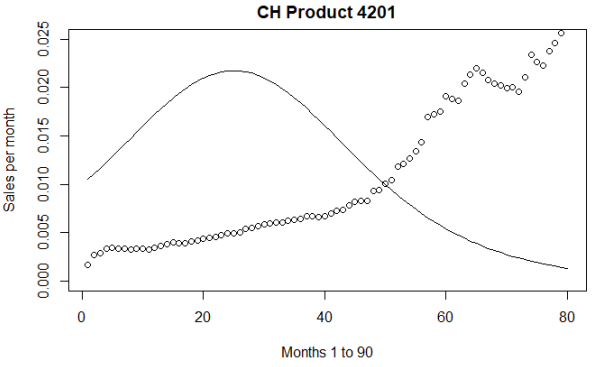
(a) M10 fit for China Product 3104



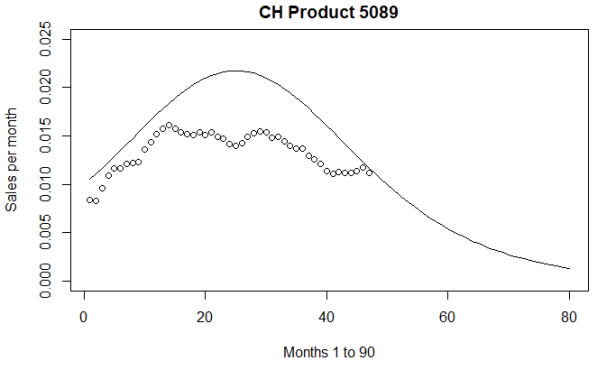
(b) M10 fit for China Product 3105



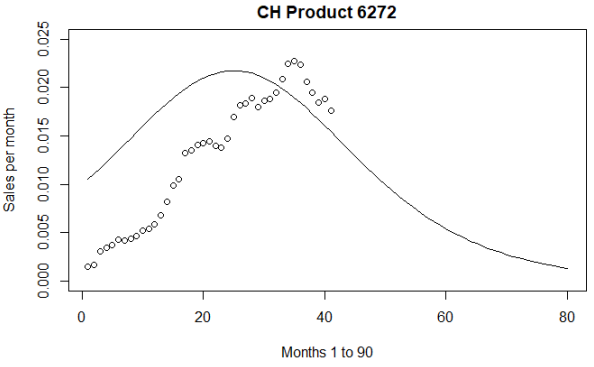
(c) M10 fit for China Product 3128



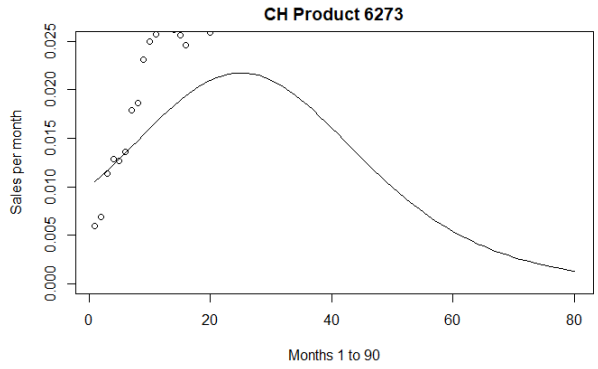
(d) M10 fit for China Product 4201



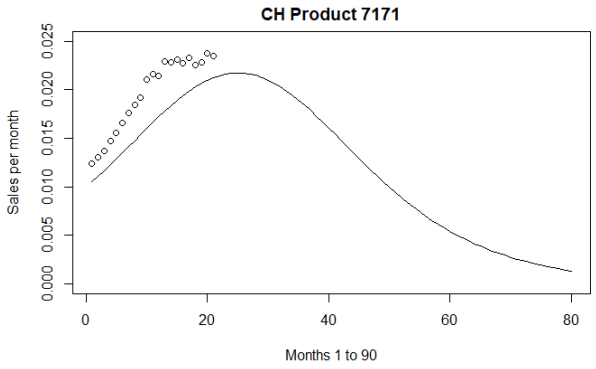
(e) M10 fit for China Product 5089



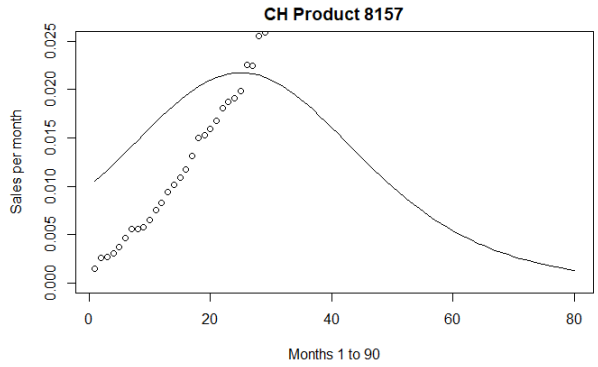
(f) M10 fit for China Product 6272



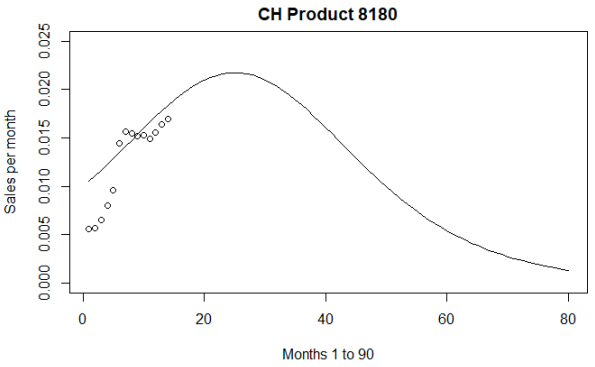
(g) M10 fit for China Product 6273



(h) M10 fit for China Product 7171



(i) M10 fit for China Product 8157



(j) M10 fit for China Product 8180

Figure 29: Global Total Sales Model (M10) fit in the China market.

References

Bass, F. (1969). A new product growth model for consumer durables. *Management Science*, *15*(5), 215–227.

Bass, F. M., Trichy, V., & Dipak, C. J. (1994). Why the bass model fits without decision variables. *Marketing Science*, *13*(3), 203-223.

Blackman, A. w. (1972). Mathematical for trend forecasts. *Technological Forecasting and Social Change*, *3*, 441-452.

Booz, Allen, & Hamilton. (1982). New product management for the 1980s. *New York: Booz, Allen and Hamilton, Inc.*

Cao, H., & Folan, P. (2012). Product life cycle: The evolution of a paradigm and literature review from 1950–2009. *Production Planning and Control*, *23*(8), 641-662.

Carden, S. (2005). What global executives think about growth and risk. *McKinsey Quarterly*, *2*, 16-25.

Chatterjee, R., & Eliashberg, J. (1989). The innovation diffusion process in a heterogeneous population: A micromodeling approach. *Management Science*, *36*(9), 1057-1079.

Coad, & Rao. (2008). Innovation and firm growth in high-tech sectors: A quantile regression approach. *Research Policy*, *37*(4), 633-648.

Cox, W. (1967). Product life cycles as marketing models. *The Journal of Business*, *40*(4), 375-384.

Decker, R., & Gnibba Yukawa, K. (2010). Sales forecasting in high technology markets: A utility based approach. *Journal of Product Innovation Management*, *27*(1), 115-129.

den Bulte., C. V. (2005). Want to know how diffusion speed varies across countries and products? try using a bass model. *PDMA visions*, *4*, 12-15.

Dodds, W. (1973). An application of the bass model in long-term new product forecasting. *Journal of Marketing Research*, *10*(3), 308-311.

Erdem, T., Keane, M., & Strebel, J. (2005). Learning about computers: An analysis of information search and technology choice. *Quantitative Marketing and Economics*, *3*(3), 207–247.

Faison, E. (1977). The neglected variety drive: a useful concept for consumer behavior. *J. Consum. Res.*, *4*, 172-175.

Fisher, & Pry. (1971). A simple substitution model of technological change. *Technological Forecasting and Social Change*, *3*, 75-88.

Floyd, A. (1962). Trend forecasting: A methodology for figure of merit j. bright (ed.), technological forecasting for industry and government: Methods and applications. *Prentice-Hall, Hinsdale, IL*, 95-105.

Fourt, L., & Woodlock, J. (1960). Early prediction of market success for new grocery products. *Journal of Marketing*.

Google Inc, a. (2018). Google scholar.

Grantham, L. M. (1997). The validity of the product life cycle in the high-tech industry. *Marketing Intelligence and Planning*, *15*(1), 4-10.

Guo, X. (2014). A novel bass-type model for product life cycle quantification using aggregate market data. *International Journal of Production Economics*, *158*, 208-216.

Jaakkola, H. (1996). Comparison and analysis of diffusion models. *Diffusion and Adoption of Information Technology.*.

Jiang, Z., Bass, F. M., & Bass, P. I. (2006). Virtual bass model and the left-hand data-truncation bias in diffusion of innovation studies. *International Journal of Research in Marketing*, *23*(1), 93-106.

Jun, D. B., Kim, S. K., Park, Y. S., Park, M. H., & Wilson, A. R. (2002). Forecasting telecommunication service subscribers in substitutive and competitive environments. *International Journal of Forecasting*, *18*(4), 561-581.

Kim, T., & Bass, F. M. (2005). A study of bias and systematic change in nonlinear estimation of bass model parameters. *ProQuest Dissertations and Theses.*.

Lattin, J., & Roberts, J. (1989). Modelling the role of risk-adjusted utility in the diffusion of innovations. *Graduate School of Business*.

Levitt, T. (1965). Exploit the product life cycle. *Harvard Business Review*, *43*(6), 81–94.

Mahajan, V., & Muller, E. (1979). Innovation diffusion and new product growth models in marketing. *Journal of Marketing*, *43*(55-68).

Mahajan, V., Muller, E., & Bass, F. (1990). New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, *54*(1), 1-26.

Mahajan, V., Muller, E., & Bass, F. M. (1993). New-product diffusion models, handbooks in operations research and management science. *Elsevier.*, *5*, 349-408.

Mahajan, V., Muller, E., & Wind, Y. (2000). New-product diffusion models. *Springer Science and Business Media*.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, *29*(4), 741-766.

Massiani, & Gohs. (2015). The choice of bass model coefficients to forecast diffusion for innovative products: An empirical investigation for new automotive technologies. *Research in Transportation Economics*, *50*, 17-28.

Orbach, Y. (2016). Parametric analysis of the bass model. *Innovative Marketing*, *12*, 29-40.

Oren, S., & Schwartz, R. (1988). Diffusion of new products in risk-sensitive markets. *Forecasting*, *7*, 273-287.

R Core Team, a. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Rink, D., & Swan, J. (1979). Product life cycle research: a literature review. *Journal of Business Research*, *7*, 219-242.

Rogers, E. (1995). Diffusion of innovations. *New York: The Free Press.*, *4*.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Science, M. (2004). Bass model. *Management Science*, *50*(12), 1833-1840.

Sharif, N., & Kabir. (1976). A generalized model for forecasting technological substitution. *Technological Forecasting and Social Change*, *8*(4), 353-364.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, *5*, 13–22.

Srinivasan, V., & Mason, C. (1986). Nonlinear least squares estimation of new product diffusion models. *Marketing Science*, *5*(2), 169-178.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*, *63*(2), 411-423.

Urban, G., & Hauser, J. (1993). Design and marketing of new products. *Englewood Cliffs, NJ: Prentice Hall.*.

Werker, C. (2003). Innovation, market performance, and competition: Lessons from a product life cycle model.

Wright, M., Upritchard, C., & Lewis, T. (1997). A validation of the bass new product diffusion model in new zealand. *Marketing Bulletin*, *8*, 15–29.

Zambelli, A. (2016). A data-driven approach to estimating the number of clusters in hierarchical clustering.

Zhengrui, J., Bass, F. M., & Bass, I. (2005). Virtual bass model and the left-hand data truncation bias in product diffusion studies. *Forthcoming Intern. J. Res. Marketing.*.