Forecasting Behavior with Age-Period-Cohort Models: How APC Predicted the US Mortgage Crisis, but Also Does So Much More

Joseph L. Breeden, Prescient Models LLC

March 14, 2016

ABSTRACT

Age-Period-Cohort (APC) models have been workhorse algorithms in epidemiology an demography for decades, but their applicability goes much wider. Recent successes in retail loan stress testing and modeling through the US mortgage crisis have highlighted how these algorithm can be used much more broadly to obtain unique insights.

This paper provides an introduction to APC models, which analyze data in which performance is measured by age of an account, account open date, and performance date. We also provide details of their implementation and limitations.

This flexible technique is demonstrated with an example from a recent study that reveals root causes of the US mortgage crisis not observed from conventional discussions. In addition, we include demonstrations of the application of APC models to forecasting the value of fine wines, member lifetime value to the SETI@home project, and tree ring analysis for climate change. Lastly, we discuss how APC models can predict website usage, retail store sales, salesperson performance, and employee attrition. We even present an example in which APC was applied to a database of tree rings to reveal climate variation in the southwestern United States.

INTRODUCTION TO AGE-PERIOD-COHORT MODELS

Age-Period-Cohort (APC) models are nothing new. They evolved from the Lexis diagrams created in the 1890s as a way to view mortality trends for different cohorts. In mortality, a cohort is just defined as the year of one's birth. In the 1960s and 70s, techniques were developed to analyze aggregate time series of cohort performance in order to measure the main drivers of performance.

Although begun with mortality studies, APC models became most popular in studies of epidemiology and demography. This paper seeks to introduce APC models to a broader audience and demonstrate the breadth of applications possible through specific examples from banking, fine wines, SETI@home, and tree rings. In addition, we describe how they could be applied to many other areas.

In the case of mortality modeling, the probability of death should follow a binomial distribution, so an APC model could be defined as

$$log\left(\frac{p(a,p,c)}{1-p(a,p,c)}\right) = F(a) + G(c) + H(p)$$

where the left side of the expression is the logit transformation of the probability of death $p(a,c,p)$, a is the age at death, c is the birth cohort, and p is the period (calendar date) at death. Three functions are estimated, one each for age, cohort, and period, $F(a)$, $G(c)$, and $H(p)$ respectively. Note also that $a = p - c$, the age of the entity can be computed by subtracting the cohort date from the period.

The reason is that these functions are estimated without explanatory factors. In cases where the root causes are unknown, the model can still measure the impact of the "environment" on a given calendar date, variation by cohort, and timing versus age.

Stated more generally, APC models model some rate by estimating functions of age, cohort, and time with a link function appropriate for the distribution of the event. In other areas of application, the terms age a, vintage v, and time t are used in place of age, cohort, and period.

$$r(a, v, t) = Link\big(F(a) + G(v) + H(t)\big)$$

Intuitively, the corresponding functions as estimated for some rate can be described as:

Lifecycle, $F(a)$: This quantifies the expected rate as a function of age at time of event. The magnitude is scaled to the rate being modeled. In later examples, this could be the probability of default on a loan versus the age of the loan or the probability of making an online purchase versus the age of the customer account.

Vintage quality, $G(v)$: A measure of variation by the origination date (a.k.a. cohort or vintage date) for a group of people, customers, loans, etc. For a consumer loan, it would measure the net credit risk by vintage after normalizing for lifecycle and environment. For online sales, it could be relative propensity to purchase versus the date or even time when the customer first registered. The magnitude is measured in terms of the relative change in the rate where the zero level is the average and the vintage estimates may be roughly normally distributed about that.

Environment, $H(t)$: A measure of the net impact of all factors that impact performance as a function of calendar date. In most business applications, one thinks of this including macroeconomic drivers and any management policy changes that affect all customers simultaneously. The magnitude is again scaled as the relative change in rate with zero representing the average.

SIMILARITY TO OTHER TECHNIQUES

The basic concept of APC models is similar to survival and hazard models. In all these methods, an entity is being monitored for the probability of some event. The age of the entity is a key determinant of the event rate. For example, part failures tend to increase versus the age of the part. This is referred to as the hazard or lifecycle function.

Survival models measure the survival function as the probability that an entity will survive to a certain age. The hazard function is the probability that an event will occur at a given age conditioned on the event not having already occurred. Thus, survival models capture one of the three dimensions (age) of the APC models. Survival or hazard functions are generally estimated nonparametrically. In SAS this can be done with

```
proc lifetest
```

Cox Proportional Hazard models (Cox PH) are an extension to survival models that include factors that adjust the probability of the event. The usual notation is

$$\lambda(a|\boldsymbol{X_i}) = \lambda_0(a)exp(\boldsymbol{X_i} \cdot \boldsymbol{\beta})$$

where $\lambda(t|\boldsymbol{X_i})$ is the event rate conditioned on scaling factors $\boldsymbol{X_i}$ specific to entity i and $\boldsymbol{\beta}$ are the coefficients to be estimated. David Cox showed that the estimation of the coefficients is independent of the estimation of the hazard function, so he developed a partial likelihood optimization procedure for estimating the coefficients. In SAS this is found in

```
proc phreg
```

The Cox PH coefficients could be used to estimate the vintage or environment function parameters, but are more commonly applied to explanatory factors. The challenge is that Cox PH was not developed with the three dimensions of age, vintage, and time in mind, where a linear relationship exists between these dimensions. Therefore, the APC framework gives us more explicit control over the estimation of these functions, which becomes of critical important during forecasting.

MODEL ESTIMATION

Although APC models are publicly available, there are some choices to be made when using them.

DATA STRUCTURE

Because APC models estimate their functions just from information about the time of the event, the data structure is extremely simple. For example, the following dataset would be sufficient for most APC estimations. The data snippet in Table 1 could be used to estimate a model of

$$Attrition\ Rate(t) = \frac{Closed\ Accounts(t)}{Active\ Accounts(t-1)}$$

From the vintage and performance dates, the APC algorithm computes the age and estimates all three functions.

Vintage Date (Cohort)	Performance Date (Period)	Active Accounts (t-1)	Closed Accounts (t)
Jan 2010	Feb 2010	100	2
Jan 2010	Mar 2010	98	3
…			
Feb 2010	Mar 2010	120	1
Feb 2010	Apr 2010	119	0
Feb 2010	May 2010	119	2
…			

Table 1. Data structure for an APC model.

Model Estimation

Before estimating the three APC functions, we must again consider the relationship $a = t - v$, age equals time minus vintage. If we start by assuming that the three functions of age, vintage, and time are completely general, then they must have constant, linear, and nonlinear parts. Without losing any generality, this can be expressed as

$$F(a) = \alpha_0 + \alpha_1 a + F'(a)$$
$$G(v) = \beta_0 + \beta_1 v + G'(v)$$
$$H(t) = \gamma_0 + \gamma_1 t + H'(t)$$

Substituting into our APC equation, we get

$$p(a, v, t) \sim \alpha_0 + \alpha_1 a + F'(a) + \beta_0 + \beta_1 v + G'(v) + \gamma_0 + \gamma_1 v + H'(t)$$

This equation has a few obvious problems. α_0, β_0, and γ_0 are the constant terms for the three respective functions, however, we cannot simultaneously estimate three constant terms in the same equation. Therefore we define a single constant term $\alpha_0' = \alpha_0 + \beta_0 + \gamma_0$. This means that the constant term is being included in the age function and all other functions are estimated relative to the age function. This is equivalent to what is done in Cox PH where the hazard function is scaled to the historic probability and the Cox PH regression coefficients are measured relative to it.

In addition, because $a = t - v$, α_1, β_1, and γ_1 cannot all be estimated independently. Only two of the three linear trends can be estimated independently. To resolve this, an assumption must be made about the allocation of the linear components among the three functions. In applications where the rate being modeled is stationary through time, we may assume that $\gamma_1 = 0$ so that the other two linear terms can be estimated uniquely.

Holford (1983) explains well that a decision must be made about the trend ambiguity in any vintage analysis, whereby only two linear components can be estimated from the three available dimensions. This decision is always domain-specific. Breeden, Bellotti, and Yablonski (2015) also explained that this problem occurs for any model of vintage data. Methods like Cox PH provide unique solutions only because of trend allocation assumptions embedded within the estimator. Breeden and Thomas (2016) provide specific solutions for models that incorporate macroeconomic factors.

Holford also proves that aside from the constant and linear term considerations, all other components of the functions are uniquely estimable. Therefore, the primary issue in using APC models is to determine the linear trend allocation. For the analyses shown here, sufficient history is available for us to assume that the environment function has no linear trend. Therefore, the following assignments are made for the functions.

$$F(a) = c_0 + c_1 a + F'(a)$$
$$G(v) = c_2 v + G'(v)$$
$$H(t) = H'(t)$$

ESTIMATORS

APC models come in many flavors (Yang and Land, 2013). In all implementations, the functions are estimated without reliance on external explanatory factors. Rather, the functions are estimated either parametrically via splines or some other basis functions, or nonparametrically.

In SAS, APC models may be estimated using the following procedures

- Spline estimation is available via `proc transreg`.
- Bayesian estimation (Schmid and Held, 2007) is available via `proc genmod`.
- Partial least squares estimation is available via `proc pls`.
- Ridge regression estimation is available via `proc reg (w/ridge= option).`

These techniques are compared for a sample dataset in Figure 1.

Spline Estimation

The most common implementations use spline estimation to approximate the three functions, but several nonparametric estimation techniques exist as well. Splines are piecewise polynomial function approximations, where the user must specify the number and potentially location of the spline nodes. In the author's experience, fewer nodes are effective for the age function, since the prior assumption is usually that it should change smoothly with age. For the time and vintage functions, the nodes are best distributed according to the density of the data, and with as many nodes a supportable by the data. Both of these functions can have frequent discontinuities, because of sudden changes in the environment or in origination conditions for the vintages.

Bayesian Estimation

A Bayesian estimation algorithm was proposed by Schmid and Held (2007). Their algorithm estimates the functions nonparametrically, meaning that each point of the lifecycle, environment, and vintage functions is a separate parameter estimated without the smoothness constraint of the spline algorithm. It can use an initial prior for those functions, but then uses a Monte Carlo estimation procedure to refine the functions. Because the Monte Carlo simulation can be very slow, a quick spline estimation can serve as a useful prior to accelerate convergence.

The Bayesian result looks “noisier”, but as a nonparametric estimation it has the ability to better estimate the complex structure that can arise in the functions. For example, sudden shocks may occur in the environment function due to operational failures. Similarly, the vintage function may have seeming discontinuities because of changes in underwriting policies.

For Bayesian estimation, the essential parameters are the number of Monte Carlo simulations used to create the estimate and the step size for each simulation.

Partial Least Squares

As an alternative to Bayesian estimation, one can also use partial least squares (PLS). PLS creates a set of principle components, or eigenvectors, on which the regression is performed. The primary control is in the number of components to be used in the estimation.

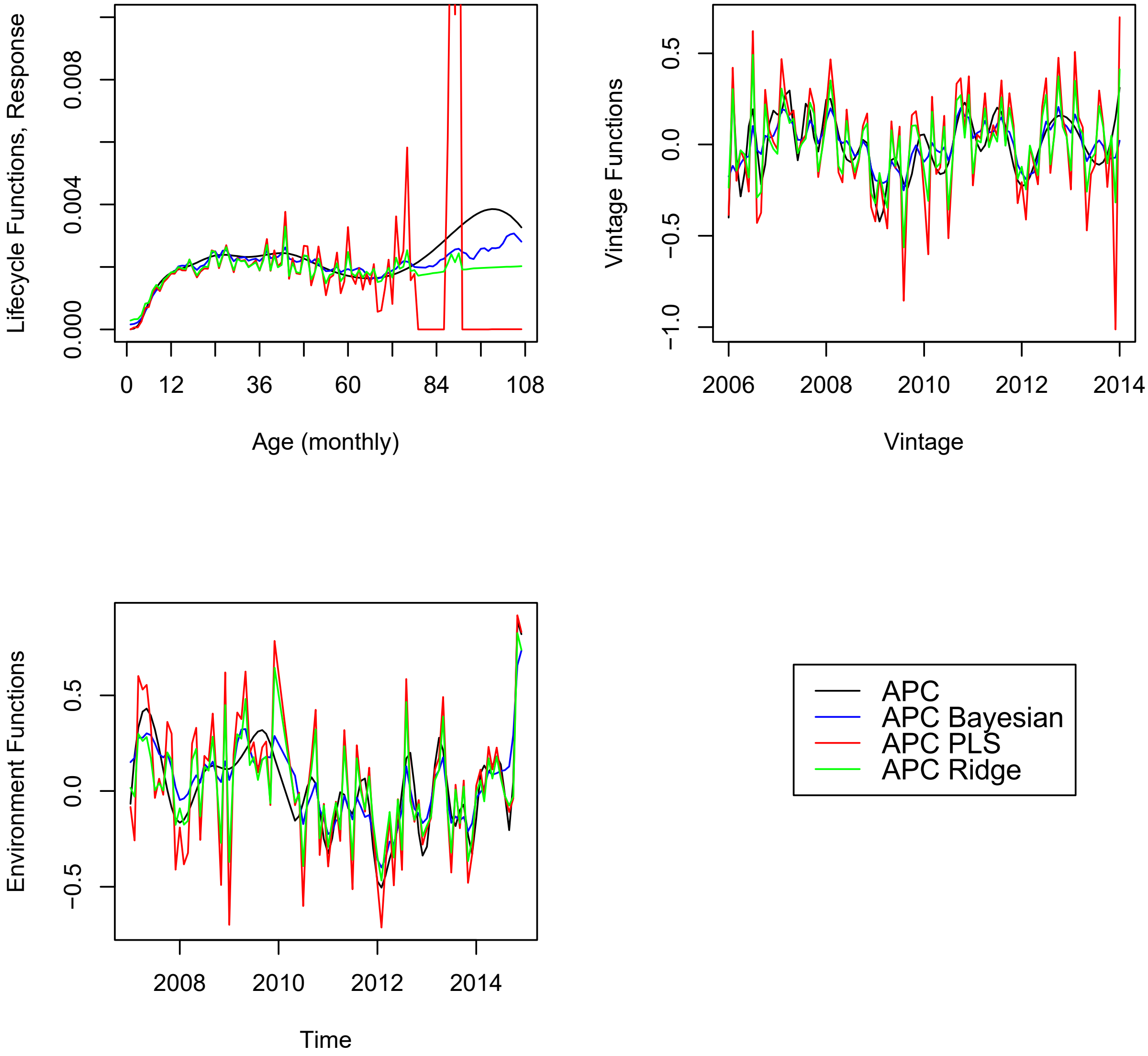

Figure 1: A comparison of APC estimation techniques as applied to modeling probability of default (PD) for a $400 million auto loan portfolio.

Ridge Regression

Ridge regression is the last technique we have used for APC estimation. It can also be an effective estimation method, although on some data sets it may have convergence problems depending on the step size used for the optimization.

In the above example, the different estimation techniques are compared. This example suggests that one could rank the smoothness of the result as APC (spline), Bayesian, Ridge regression, and PLS. However, on other data sets, we have observed a reordering of these.

One important point of comparison is how they perform for recent vintages with few observations. Spline estimation is generally unstable for these vintages whereas Bayesian estimation can use a starting prior of an average vintage and slowly diverge as observations accrue.

CROSS-TERMS AND SEGMENTATION

APC models, or vintage models more generally, assume that age, vintage, and time are independent. Independence means that there are no cross-terms in the model between age, vintage and time. This means that young and old accounts respond the same in percentage terms to changes in the environment, and that new vintages and old vintages follow the same age function.

Intuitively, however, there may be situations where cross-terms exist. In an auto loan portfolio, the new vintages might have a higher percentage of 7-year term loans and those loans have a longer lifecycle than the dominant 5-year loans. For online shoppers, those who become new users in December may just be Christmas shoppers, so their repeat-sales lifecycle may be different from those who first sign up at other times of year.

Methods exist for testing for the presence of cross-terms. Breeden (2010) demonstrated a technique that is basically a spatial correlation test on the residuals in the dimensions of age and time. The situations hypothesized above would create concentrations or positive or negative errors in specific regions of the age-time space.

Whenever cross-terms are found to be present, the solution is almost always to create separate segments for modeling. In the examples just given, auto loans should be segmented by term of loan, because different terms have different lifecycles. Once separate APC models are created by term, the analysis becomes stable. Online shoppers can be segmented between Christmas and regular shoppers to improve the modeling accuracy of both groups.

The only case the author has seen so far where the cross-terms could not be resolved by segmentation are in modeling attrition on fixed rate loans. The lifecycle function is always present, capturing the effect that consumers do not like to refinance their loans frequently. However, rather than having functions for vintage and time to capture the probability of attrition, the second strongest predictive factor is the difference between the current loan refinance rate and the original interest rate on the loan. That difference between the interest rate on calendar date and on vintage date is the definition of a cross-term. In that case, we keep the APC lifecycle, but replace vintage and environment with an interest rate factor. Such occurrences are rare.

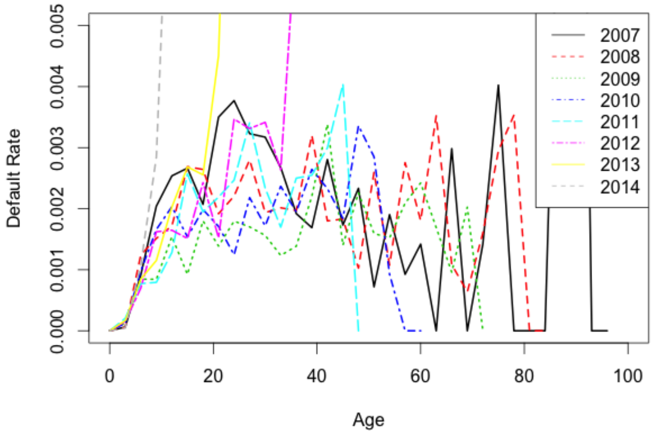

Figure 2: A plot of default rate for annual vintages aligned by age of the loan in months.

APPLICATIONS OF APC

Because the technique is so intuitive, APC algorithms have been developed independently in many fields. In retail lending, vintage analysis has been around for decades. This is the practice of making comparative graphs of loan performance in order to understand relative credit risk between pools. Decomposition into age, vintage, and time functions was developed by Breeden (2007) and can be seen as a variation on APC models.

The US Mortgage Crisis of 2009 and subsequent government stress testing requirement, Comprehensive Capital Analysis and Review (CCAR), have highlighted the usefulness of APC modeling for forecasting and stress testing. As one of the leading methods in this space, performance measures such as loan delinquency rate are decomposed to show the expected delinquency versus the age of the loan, the credit risk by vintage, and the impact of the environment on the portfolio. Secondary modeling is often performed with macroeconomic factors to explain the environment function (Breeden, Thomas, and McDonald, 2008) or scoring factors to explain the vintage function (Breeden, 2013).

In the example shown here, a US auto loan portfolio for $400 million in outstanding loans was decomposed to show the dynamics of the portfolio. Spline function estimation was performed in order to demonstrate a typical level of smoothing. The input data was numerators and denominators for the default rate data shown in **Error! Reference source not found.**.

The spline APC estimation produced the lifecycle function shown in Figure 2. In the context of retail lending, this is also referred to as the loss timing function. We can see that it captures the overall shape of the vintage data in Figure 2. Because industry intuition is that not expect much detail structure should exist, relatively few spline nodes were used.

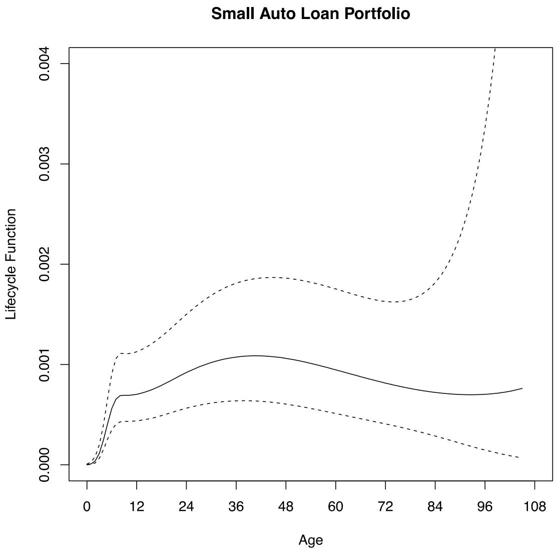

Figure 3 The lifecycle function estimated from the data in Figure 2.

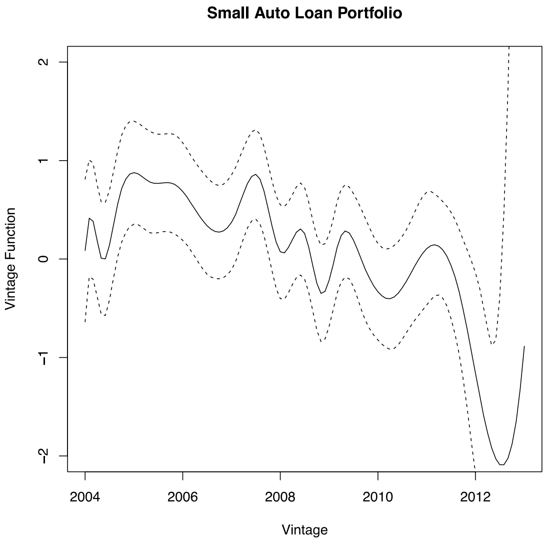

Figure 4: The vintage function showing the credit risk by vintage. The zero level means the same as the lifecycle. Higher numbers correspond to more risk.

Figure 3 shows the vintage function estimated. In this case, more spline nodes were used in order to provide greater detail on how underwriting changed through time. Note in this graph that loans originated before 2008 are higher risk (almost 1.0 on the scale). After the mortgage crisis, underwriting was tightened so the loans in 2010 and after are below 0.

Figure 4 shows the environment function versus calendar date. The 2009 recession is clearly visible as a period of higher-than-normal losses. The 2011 recovery is also clear. The rise at the end of period is due to a change in default recognition policy. This ability to capture non-economic portfolio impacts is one of APC's unique advantages in application to retain lending.

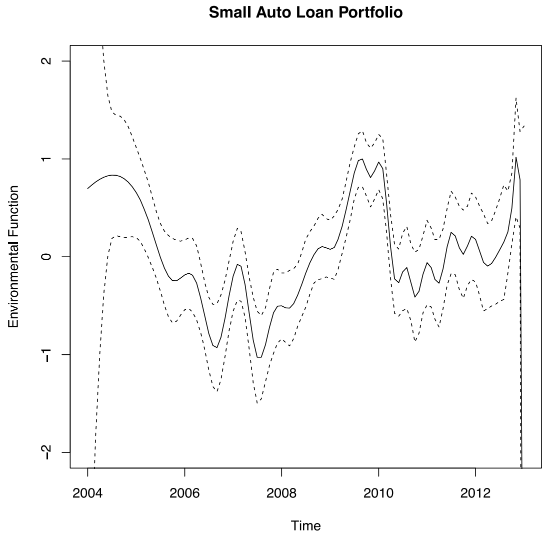

Figure 5: The environment function versus calendar date. The zero level is the historic average. Positive numbers mean higher-than-average losses.

Note that for the most recent vintages, the uncertainty grows dramatically, because very few observations are available. The same is true for the oldest part of the lifecycle function and the oldest part of the environment function.

By creating scenarios for the future of the environment function, usually via regression models with macroeconomic data, forecasts and stress tests can be created (Breeden, Thomas, and McDonald, 2008). To predict the performance of future vintages, a scenario is required for the quality of those originations on the same scale as the vintage function. In both cases, these scenarios are combined with the known lifecycle function in order to create forecasts.

THE US MORTGAGE CRISIS

The preceding analysis was replicated on an industry-wide data set for prime 30-year fixed rate confirming mortgages by Breeden and Canals-Cerda (2016), but with an additional step to include factors known about the loans at time of origination. The age, vintage, and time functions were quantified via APC, but rather than just stop with a vintage function capturing relative risk of the vintages, a second model was created to include explanatory factors.

$$log\left(\frac{p_i(a,v,t)}{1-p_i(a,v,t)}\right) = F(a) + H(t) + c_0 + \sum_{j=1}^{n_s} c_j s_{ij} + \sum_{v=1}^{n_v} g_v$$

In the above equation, $F(a)$ and $H(t)$ are both fixed inputs obtained from the APC decomposition. The coefficients c_i are estimated for explanatory scoring factors s_{ij} for account i. The vintage fixed effects (dummy variables) g_v are similar to the original vintage function, but now they will capture only the residual structure not explained by the scoring factors.

For the mortgage dataset, scoring factors were available for jumbo loan (y/n), documentation level, FICO at origination, loan-to-value at origination, loan source, occupancy, PMI (mortgage insurance; y/n), term, and purpose. Given the large size of the dataset, all of these factors were significant.

The original APC analysis by Breeden and Canals-Cerda showed that the vintages originated prior to the recession were significantly more risky than the best (lowest risk) vintages of 2003 and 2009. Comparing the worst vintages of 2007 and 2008 to the best vintages of 2003 and 2009, the change in log-odds of delinquency was 2.0, roughly equivalent to 100x greater delinquency comparing the worst loans to the best. Note again that this difference was independent of the recession, as that effect is captured in the environment function $H(t)$.

The goal of the analysis was to determine if all of the increased credit risk by vintage was due to known underwriting factors or other causes. In fact, Figure 6 shows a comparison of the APC vintage function just described to the residual vintage fixed effects, g_v, after normalizing for lifecycle, environment, and available scoring factors. From the graph, we see that only half of the originally estimated APC vintage function can be explained by known underwriting factors. The rest has some other cause.

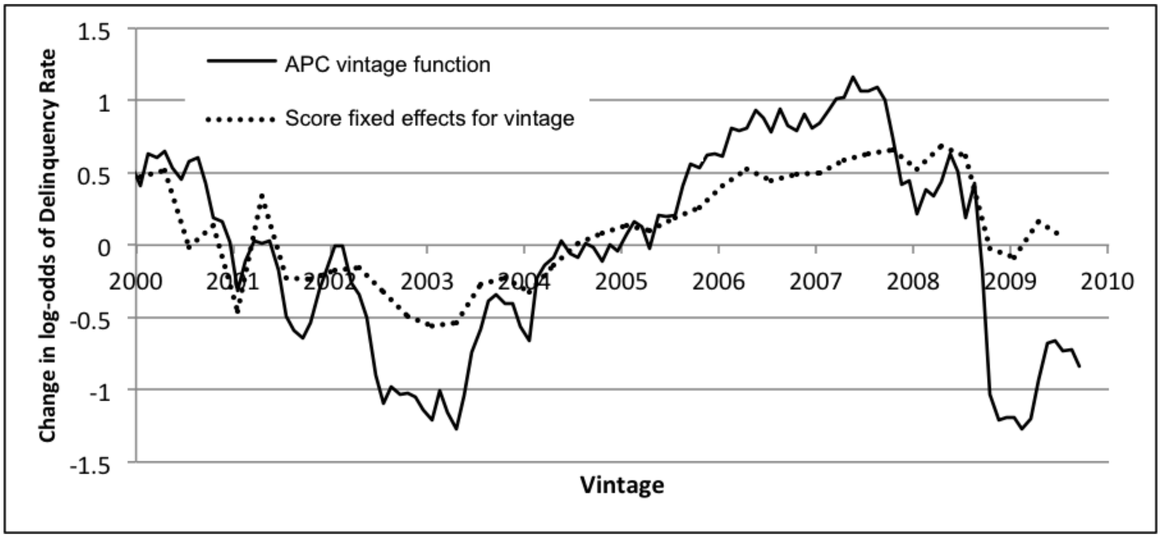

Figure 6: A comparison of the vintage function from APC decomposition and the score fixed effects for vintage capturing the unexplained residual after including explanatory scoring factors (Breeden & Canals-Cerda, 2016).

Subsequent analysis by Breeden and Canals-Cerda showed that the residual vintage fixed effects correlated well to the 2-year change in mortgage interest rates at the time the loan was originated. The implication is that credit risk is driven by more than just the underwriting decisions of the banks. Consumers are looking at interest rates and economic conditions to decide if it is a good time to get a loan.

The APC algorithm allowed the analysis to normalize for lifecycle and environmental effects. In addition, this analysis showed how to combine APC decomposition with traditional scoring.

WINE PRICE FORECASTING

With any APC analysis, the first question is how to define a vintage. Fine wines must be the most obvious example possible. The wine’s vintage is the year the grapes were harvested. Although the bottled product is not sold immediately, we count the age of the vintage from the harvest year.

For the analysis, a database was provided by auctionforecast.com covering fine wine auction prices over a 15 year time span from the following auction houses: Acker Wines, Bonhams, Chicago Wine Co., Christie's, Langton's, Sotheby's, Spectrum Wine, Veiling Sylvie's, and Zachy's. The provided data adjusted all currencies to US dollars according to the exchange rate at the date of the auction. All prices are hammer prices (before expenses) without adjustment for inflation. Only auction results for homogenous lots were included (meaning the auction price was only for bottles of the same wine and vintage), and all prices were converted to price per bottle. Although some information is available in the lot descriptions for items like ``original wooden case" or ``damaged label", these were not included in the modeling database.

For the subsequent analysis, only wines that had been sold at auction at least 16 times were included. A ``wine" is defined as a specific make of a specific vintage. Although the database contains thousands of wines, the example here focuses just on the make most traded at auction, Château Lafite Rothschild.

Since the auction prices follow a lognormal distribution, the following for was used for APC analysis.

$$\log\big(price(a,v,t)\big) = F(a) + G(v) + H(t)$$

When applied to auction price data for fine wines, the lifecycle function, F(a), measures the expected average price for a wine in a segment as a function of the age of the wine. Thus, the lifecycle shows the expected rate of appreciation in a wine’s value across different spans of age.

The vintage function, G(v), captures how much higher or lower a given wine is priced relative to the average lifecycle for the segment. This allows for the estimation of separate price scaling by vintage while maintaining a common market index (environment function) and common lifecycle function across all Lafite wines.

The environment function, H(t), measures how much auction prices are above or below the expected lifecycle values on a given calendar date.

Lifecycle

The lifecycle estimation for price versus age of the wine was first done with a Bayesian APC estimator (Figure 7A). The lifecycle function was then smoothed for business purposes, but the confidence intervals are preserved.

The lifecycle shows that the average auction price actually declines until the 5th or 6th year, at which point the prices stabilize and begin to rise again. The same effect is seen for all wines at auction with some variation in where the bottom occurs.

The most rapid price increases occur in the couple decades after that minimum before slowing their rate of appreciation throughout the remaining lives of the wines. The cumulative price appreciation between 5 and 25 years of age is 81% or 3.2% annually. The importance of these estimates is that they are cleaned of changes in market conditions and represent the performance of average Lafite wines cleaned of differences in specific vintage performance. Although 81% appreciation sounds impressive, 3.2% annual capital appreciation for the wine is less exciting when considering transaction costs, storage costs, inflation, etc.

Vintage

The vintage function (Figure 7B) captures the overall trend in increasing prices for newer Lafite vintages, and the exceptional prices for some vintages. The highest spikes in the graph correspond to the 1982, 2000, 2009 and 2010 vintages, which aligns well with industry expertise.

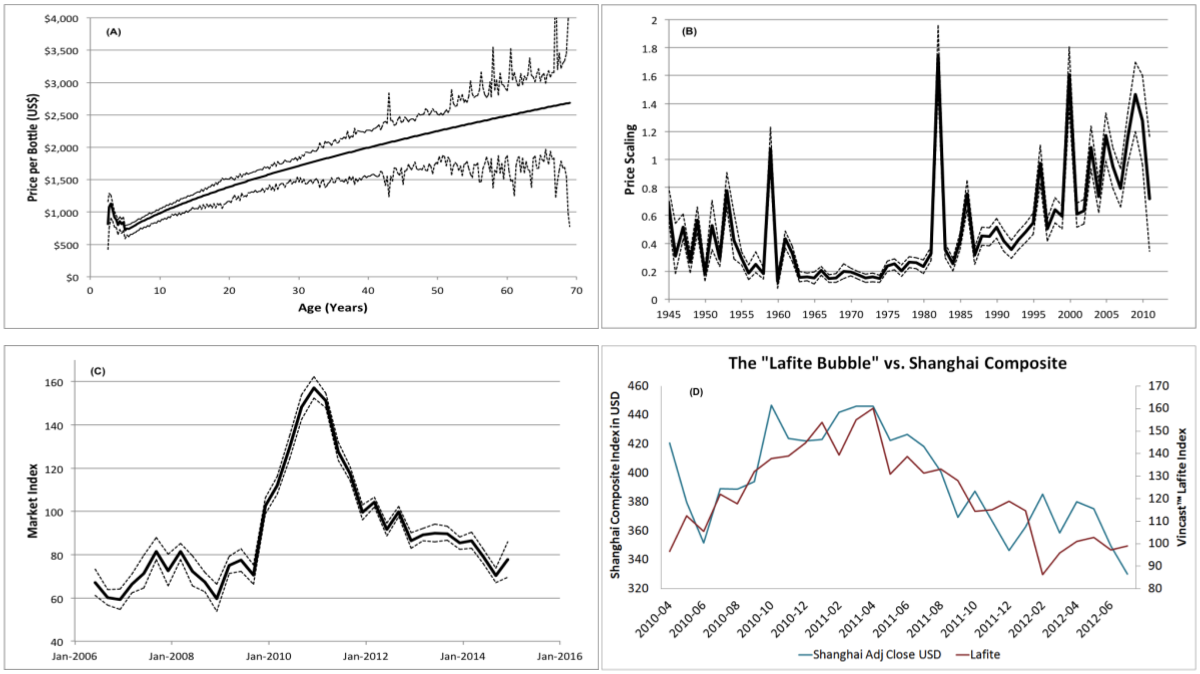

Figure 7: The APC decomposition for Château Lafite Rothschild wines: (A) lifecycle, (B) vintage, and (C) environment. Then an overlay of the environment versus the Shanghai Composite stock index

Environment

The environment function measures in the decomposition provide significant new insights. Many wine market indices are available, expressed as baskets of specific vintages. Like a stock market index, these baskets can be changed over time to swap in newer vintages, but wine-basket indices have are problematic as measures of the wine market. As seen in the lifecycle analysis, wines appreciate over time. Therefore, even if buyer interest is flat, a wine basket index will continue to rise unless manually readjusted. Further, these baskets generally have a small number of select wines and therefore do not capture the broader market conditions. Therefore, when annualized returns are quoted for specific wine portfolios, we cannot immediately conclude the causes of the appreciation, whether inherent to the wine or due to market trends.

A market index (Vincast Lafite Index, VLI) was created from the environment function for ease of interpretation by financial analysts as

$$VLI(t) = \exp\big(H(t)\big) * 100$$

A value of 100 represents the historic average for the market. The traditional approach to gauging the wine market is to create a basket of top wines and track their value. However, as the analysis here shows, that approach confounds the normal appreciation from the lifecycle with changes in market conditions. APC provides a market index that can leverage all wines auctioned, not just those in a select list, and is normalized for the natural appreciation in the value of the wines over time.

In contrast, the environmental functions in Figure 7C are also market indices, but with broad coverage (all vintages with a minimum number of auction results, set to 16 for robustness) and normalized for lifecycle and vintage effects. Normalizing by lifecycle means that the appreciation discussed in the previous section, including inflation is removed automatically. Normalization by vintage means that prices for highly valuable wines are adjusted to a measurement of changes that is comparable to price movements in less valuable vintages.

The market index for Lafite clearly shows the peak in early 2011 known as the “Lafite Bubble”. From June 2010 to Feb 2011, prices for Lafite wines, adjusted for lifecycle and vintage effects, jumped by roughly 50%. By April 2013 the environment function shows a decline to levels below the June 2010 start of the bubble.

Conventional wisdom is that the Lafite Bubble was caused by a sudden increase in interest from Chinese investors. Interestingly, the Lafite environment function is highly correlated to the Shanghai stock market index throughout this same timeframe, Figure 7D. This finding of correlation is different from that of some previous studies over different time periods, but the different results may reflect a change in the wine market.

Overall, analysis by Vincast shows that the Lafite, Bordeaux excluding Lafite, and Burgundy environment functions show significant correlation to the Hong Kong and Shanghai stock market indices from Jan 2006 through May 2014, offering some evidence that Chinese wealth may have been driving the fine wine market, confirming with comments from market watchers. However, since the end of 2012, all wine environment functions have exhibited steady, significant declines regardless of stock market movements. Fine wines appear to be in a bear market through 2015.

SETI@HOME

SETI@home, an offshoot of the Search for Extraterrestrial Intelligence (SETI), is a scientific non-profit organization housed at the UC Berkeley Space Sciences Laboratory utilizing donated computing cycles from millions of users worldwide in the search for evidence of extraterrestrial life by combing massive amounts of radio telescope data for signs of structured communication.

The project was initiated in 1999 and has grown to become the largest distributed computing project in history. SETI@home users download a screensaver program that harnesses idle time on the user's computer to mine data packets shipped automatically over the Internet. Over 600,000 CPU years had gone into the effort in just the first two years, making it the largest computer processing capability in the world. The success of SETI@home has spawned a number of other efforts in areas like protein folding, cancer research and other scientific computing problems.

Because SETI@home is a voluntary undertaking, it is subject to the same management complexity present in consumer financial products with variable usage, e.g. credit cards. For SETI@home, server failures, CPU upgrades, and software upgrades have dramatic effects, whereas retail bank customers are analogously impacted by economics, competition, and policy changes. In all metrics, natural lifecycles and variations in user quality provide the same kinds of dynamics observed in other consumer contexts.

PROJECT DESIGN

The following results are derived from 165,000 accounts sampled randomly from the available 3 million users. Those accounts were grouped into vintages according to the month in which the user downloaded the SETI@home software. A 26-month period was analyzed for this study.

A new user for SETI@home begins by opening an account and downloading a version of the software for their computer. The software comes for a variety of computer platforms and sometimes in both screensaver and command line (no interface) versions. Once the program is started, it requests a data block from the SETI@home server. That block is analyzed on the user’s computer. When completed, the results are sent back to the server and a new block is requested.

For our analysis, we know when the user opened their account and when they returned a result. We do not receive the scientific result or detailed user information, but we know in which country the user lives and what software version was used to analyze each result.

VARIABLE SELECTION

For SETI@home, we found three issues to be fundamental: the attrition rate of the volunteers (activity), the time needed to process a unit, and the fraction of the CPU available for processing SETI data (load). A single user may run the software on multiple computers or a multi-CPU computer, so load may exceed 1.

Rate analyzed with APC	Definition
Activity ratio	$activity(t) = \frac{active\ accounts(t)}{active\ accounts(0)}$
CPU Time per Unit growth ratio	$cputime\ per\ utni(t) = \frac{total\ cputime(t)/\ units\ returned(t)}{total\ cputime(0)/\ units\ returned(0)}$
Load ratio	$load(t) = \frac{total\ cpu\ time(t)}{available\ time\ per\ month}$ $available\ time\ per\ month = days\ in\ month * 24 * 60 * 60$

Table 2: Rates analysis with APC for SETI@home project.

Forecasts for the key rates in Table 2 are combined to create the forecasted variables in **Error! Reference source not found.**Table 3.

Forecasted Variable	Definition
Active accounts	$active(0) = open(0) * activation\ rate$ $active(t) = active(0) * activity(t)$
Total cputime per user	$total\ cputime(t) = load(t) * available\ time\ per\ month$
Units returned per active account	$units\ returned\ per\ active(t) = \frac{total\ cputime(t)}{cputime\ per\ unit(t)}$
Total units returned	$units\ returned(t) = active(t) * units\ returned\ per\ active(t)$
Expected units per initial account	$expected\ units(t) = retention\ rate(t) * units\ returned\ per\ active(t)$
Net present value of an account	$NPV = \sum_{i=0}^{n} \frac{expected\ units(t_i)}{(1+r)^i}$ where r is the discount rate

Table 3: Forecasted outputs using the key rates modeled by APC.

The above rates and forecasts were created by Breeden (2014). From that analysis, the following discussion is an excerpt of the activity rate and NPV analysis.

ACTIVITY RATE

The inverse of attrition rate, the activity ratio tells us how many of the initially active accounts are still active after some number of months. Multiplying the activity ratio by the activation rate at month 0 tells how many accounts are still active some time later. Not all the accounts initially opened will return a result. On average, between 10% and 30% of opened accounts never return a unit.

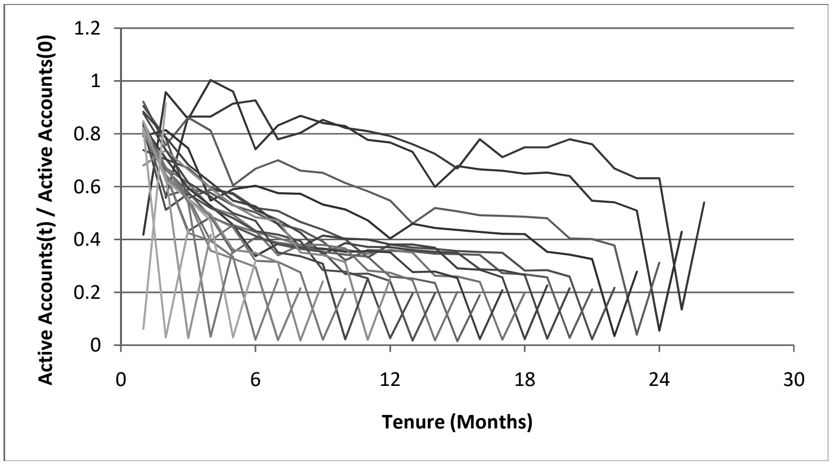

Figure 8: Raw activity ratio data, the ratio of active accounts at tenure t to active accounts at tenure 0. Each line represents a different vintage.

Figure 9 shows the decomposition results for the activity ratio obtained by analyzing the vintage curves in Figure 8. The activity lifecycle function is scaled by the initial activation rate to predict the expected activation rate. This can be read as the probability that a newly opened account will be active after a given length of time.

The lifecycle function has a prominent spike in month 1 showing that many users take more than one month to return their first unit. Thus active users surge above the month 0 level. Each month thereafter, activity falls. Most simple models use exponential rules for attrition, “x% attrite each month”. The attrition shown here actually follows a power-law structure.

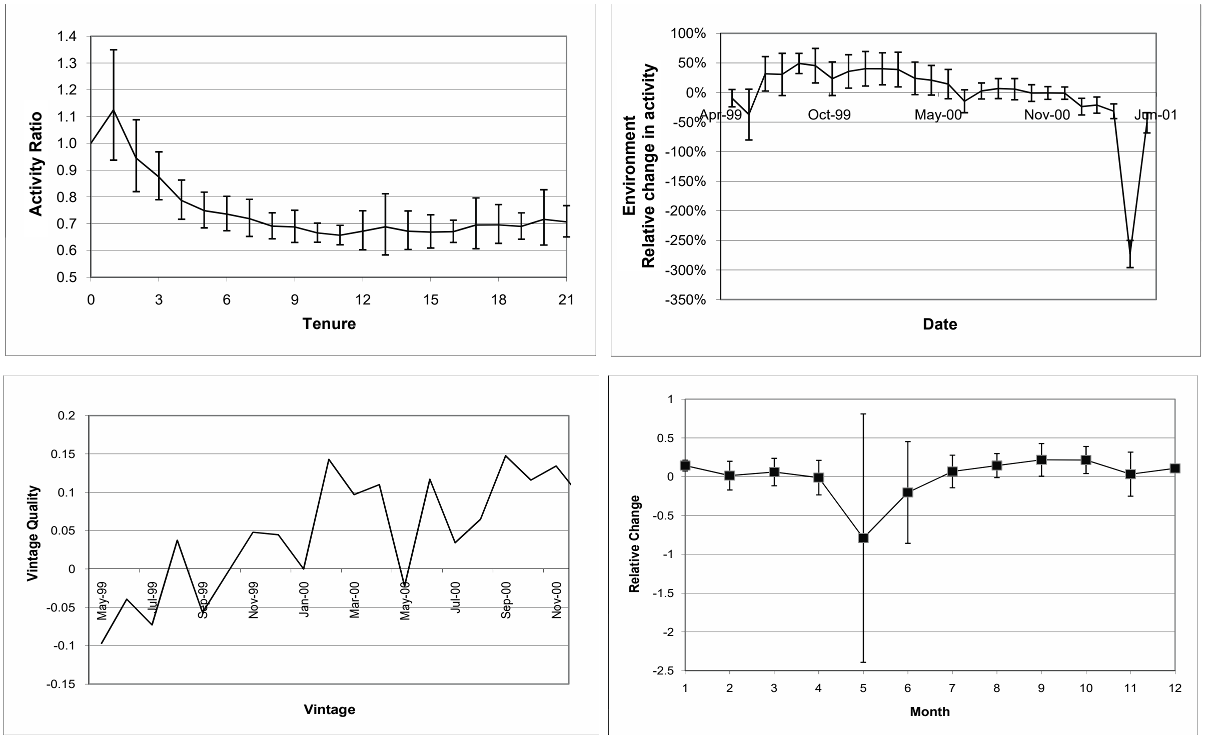

Figure 9: APC decomposition results for the activity ratio. Clockwise from upper left, these graph show the

lifecycle versus tenure (age), environment, seasonality, and vintage functions.

The environment function has obvious shocks in May 1999 and May 2001. These are easily explained by server outages that prevented users from returning results. The apparent attrition was only transient as the server was restored.

More interesting was the mandatory upgrade to version 3.03 in February 2001. Prior to this, SETI@home users were returning so many units that the network traffic was taking a major fraction of the UC Berkeley bandwidth. Version 3.03 was specifically designed to do more processing on each unit and thus reduce the bandwidth. Although the upgrade was mandatory, SETI@home is purely voluntary. As a consequence of the upgrade, many users chose not to upgrade and simply turned off the software. The lower activity level seen in February, March, and April 2001 corresponds to 19.5% more attrition than would normally have been expected. Such information is vital when planning future software upgrades.

With only two years of data, the estimate of variability based upon month of the year (seasonality) is coarse, but it does suggest certain features. Generally, activity appears to fall through June, July, and August. Activity peaks through the spring and somewhat in the fall. This structure appears to coincide with school cycles. Perhaps, many students are running SETI@home on computers at school or with their own computers on a school network. When the students leave behind either the computing or networking resources, their activity falls. A small recovery occurs when they return to school, but the time of greatest interest appears to be in the spring, when they are well settled into school.

The vintage function comprises the final component of the analysis. By examining the vintage scaling factors relative to the lifecycle, one can see that the vintage function shows a definite trend. This rise represents a stretching of the activity lifecycle function for new vintages. Since the lifecycle function decreases with time, this stretching translates to faster attrition for new vintages relative to older vintages. This might suggest that they are less dedicated to the project than older vintages.

LIFETIME VALUE

Combining the effects of initial activation, activity, time per unit, and load allows for the calculation of lifetime value for accounts. For each of these variables, the analysis was segmented by the users operating system and the country of residence.

The segmentation by OS showed many differences. The net present value calculation as shown in Figure 10 demonstrates that Windows screensaver (WinSS) users dominate the user base at around 84%, Mac users are second at 5%, Windows commandline (WinCL) and Linux users are around 2% each, and Other is >1%. Thus, most of SETI@home's processing comes from their least efficient users.

This disparity between the lifetime value of users and the concentration in the user base opens interesting possibilities. Should more effort be put into boosting activity rates for WinSS users? Should WinSS users be encouraged to move to WinCL, perhaps through repackaging the software to make the transition easy? Should Marketing and PR efforts be focused on Linux and Unix users?

SETI@home has users in almost every country on Earth, including 6 users in Antarctica. Although these users are all running the same software, local cultural and economic realities may result in different behaviors. The APC decomposition showed interesting differences by country, primarily in the rate of change in the lifecycles and the vintage functions.

Again combining load, time per unit, and activity levels into NPV, Figure 10 shows strong differences by country. Users from South Africa, Poland, Russia, the United States and Japan are the most valuable long term. The presence of the United States in the high value group is a stroke of good luck for the SETI@home project, since it provides the largest user base. If SETI@home wants to grow, however, Russia and Poland present interesting possibilities.

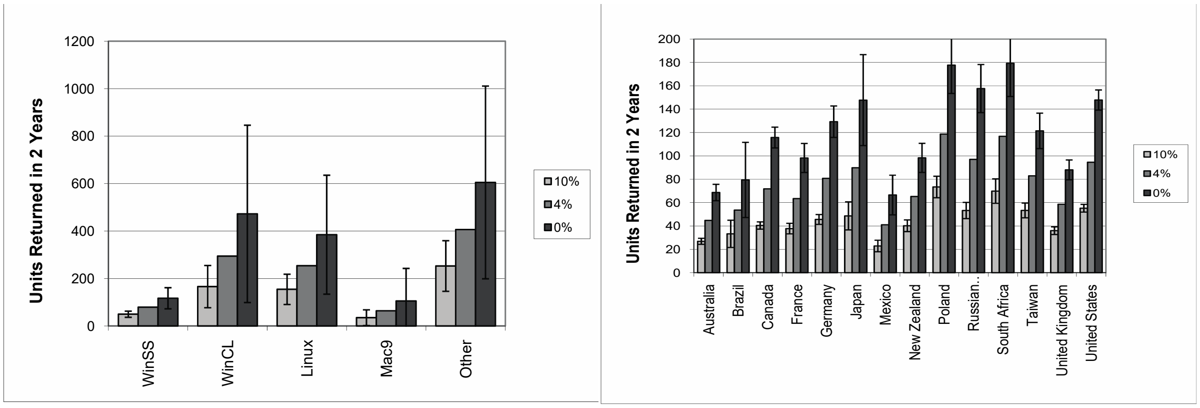

Figure 10: Net present value computed with three different annual discount rates segmented by operating system (left) and country (right).

The SETI@home project is maturing. Like any project, it has a lifecycle. Intuitively, it makes obvious sense that the earliest users were the ones most eagerly awaiting its launch. As time has passed, it is only natural that the existing users mature and new users are different from the early users. For some, enthusiasm may wane. For others, they settle into an established pattern of usage.

As the SETI@home project matures, its managers must decide what they want its future to be. If it is to grow and evolve, then, like so many other mature portfolios, more active management will be required to empower the volunteers to be as productive as possible.

DENDROCHRONOLOGY

Dendrochronology, commonly called tree-ring analysis, is an essential tool for inferring climate conditions prior to recorded history. Dendrochronology researchers use the growth patterns of trees in an attempt to infer the climactic conditions, so by definition they did not have access to external factors for the older, most interesting trees. This led them to an approach with similarities to that of the APC models.

For tree-ring analysis, each tree is treated as a “vintage” within the current framework. Unlike most consumer applications, the trees are not uniformly spaced in time. The data set has both non-uniform start and end times, i.e., trees may sprout or die at any time and must be modeled accordingly.

Although alternative measures have been explored, the standard approach is to take a core from a given tree and measure the width, w, of each growth ring. Rings occur because of seasonal dormancy periods. The amount of growth observed each year is known to be a function of the age of the tree, the environmental conditions and the specific terrain where the tree is growing. This is a perfect analogue to the lifecycle, environment and vintage functions perspective taken with retail loan decomposition.

$$\log(w) = F(a) + G(v) + H(t)$$

For this test, data was obtained from the World Data Center for Paleoclimatology. For each species of tree, all available data from the US, Canada and Mexico were used in order to estimate the lifecycle functions, one for each species of tree. Environment functions were measured independently for each site because of varying environmental factors.

In the case of dendrochronology, researchers have focused on normalizing the ring width measurements for the lifecycle effects in order to create an aggregate residual series to capture impacts from the environment (Cook and Kairiukstis 1990), the equivalent of the environment function. Researchers have no interest in unique scaling attributes of the tree, equivalent to the vintage function.

Dendrochronologists have historically removed the biologically induced growth trend (lifecycle) by fitting a mathematical function such as linear or exponential to each tree separately.

In Figure 11, 1,678 series from 59 sites were analyzed for Douglas fir, *Pseudotsuga menziesii* (Mirb.) Franco to produce a cleaned lifecycle function (left graph). From this example, we can see that the tree goes through three distinct periods. We cannot be certain that the series capture the very first years of growth, but roughly speaking, during the first 11 years of life, the growth rate is a bit less than the exponential fit shown. From ages 11 to 64, the tree fits an exponential decline in ring width very well. For ages 65 through 1000, a power law accurately describes the ring growth. The early growth divergences from the fit are due to a known growth feature of juvenile trees.

Rather than assuming a fixed functional form, some researchers have developed more general methods, such as regional curve standardization (RCS) (Briffa *et al* 1996) or age banding decomposition (ABD) (Briffa *et al* 2001). RCS is particularly interesting in the current context because trees are aligned by age to estimate a regional curve, equivalent to the maturation function in the decomposition approach. The regional curve is used to compute residuals for the individual series, which are averaged to create a mean chronology for dendroclimactic studies. ABD used somewhat similar concepts of comparing like-aged trees to obtain residual series. From this work, we see another field of research where the basic principles of APC analysis have developed independently.

The remaining tree-ring analysis focuses on data for New Mexico and Arizona: Ponderosa pine, *Pinus ponderosa* Douglas, 1929 series from 72 sites; Douglas fir, *Pseudotsuga menziesii* (Mirb.) Franco, 1678 series from 59 sites; and Piñon pine, *Pinus edulis* Engelm, 1510 series from 52 sites.

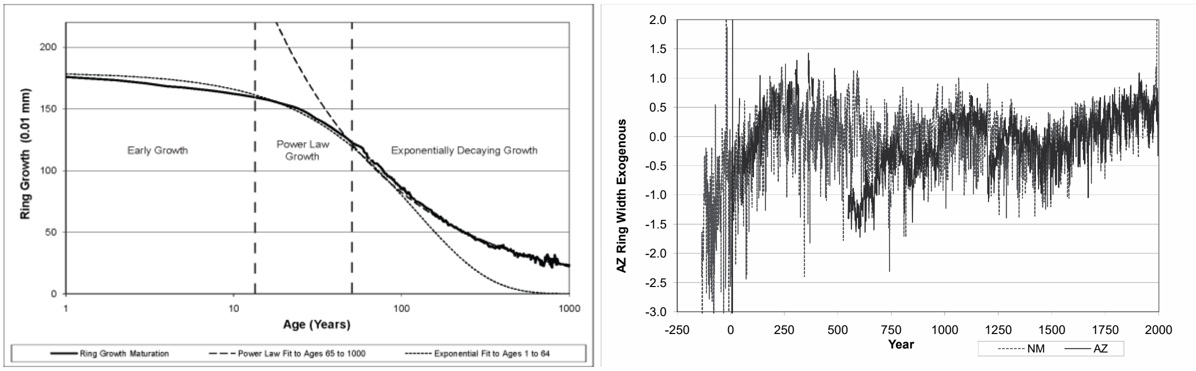

Figure 11: The APC decomposition for tree rings. At left is the lifecycle for one species showing the different growth rates versus age. At right are the environment functions for Arizona and New Mexico.

Figure 11 at right shows an overlay of the New Mexico and Arizona exogenous functions. They exhibit a remarkable amount of agreement, with the exception of the early divergences in the AZ series that appear to be due to data sparseness and integrity problems.

A 200-year rainfall series is available for New Mexico. The log of the rainfall series is shown compared to the NM exogenous function in Figure 12 at left. Visually, there appears to be a significant correlation between the two series. Using this relationship, historic rainfall levels were predicted back to 250 AD, Figure 12 at right.

The extrapolated rainfall series captures the drought of the 1950s as well as the recent wet period. These results are important to the debate over the disappearance of the Anasazi culture. Grissino-Mayer (1996) demonstrates how the booms and busts of the Anasazi civilization correspond well to the peaks and troughs of the rainfall index. The final depopulation of the Anasazi settlements occurred between AD 1300 and 1540. The causes of this disappearance are still in dispute, but the analysis here shows that rainfall levels fell dramatically during the collapse of the Anasazi civilization and did not recover until the 1600s after the colonization of the Spanish.

Although the fit of the exogenous curve to rainfall seems to agree well historically, the fit since 1984 is not as effective. During that period, the rainfall we would have expected from the environment function soared to extraordinarily high levels that were not observed. This bears further examination, but so far the analysis appears to be robust.

The environment function is not, of course, expressing rainfall alone. It is a measure of changes in tree ring growth through history that appears to have correlated well to rainfall prior to 1984. Obviously, many factors can influence the growth rate of trees. One well known environmental change through the 1900s is the increase in CO2 in the atmosphere and the ensuing debate over global climate change. Understanding the relationship between the growth of pine and juniper trees in the South-western US and atmospheric CO2 levels would be a significant research project in itself, but such a relationship is at least a plausible explanation for this divergence.

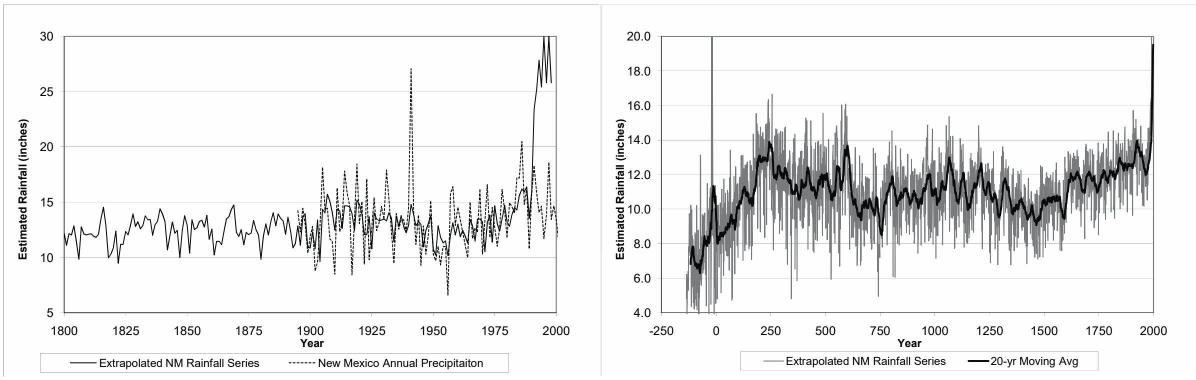

Figure 12: At left is a comparison of New Mexico tree ring environment function with New Mexico rainfall. At right is a backward extrapolation of a simple linear calibration of rainfall to the environment function.

The tree-ring analysis is interesting in that dendrochronologists have recognized the importance of “vintage” effects and have independently developed estimation methods similar to those developed in demography, epidemiology, consumer lending and elsewhere. Finding such analytical correspondences is very useful so that techniques may be shared across fields. Dendrochronology has a rich literature, which may benefit from the decomposition techniques presented here, but may also offer insights useful to refining these techniques.

OTHER APPLICATIONS

Many many other applications are possible with Age-Period-Cohort models. Due to space limitations, summaries of only a few applications have been shown, but the author has had an opportunity to apply the models to several other problems. The following list summarizes the kinds of questions that can be answered with APC models.

- **eCommerce**: Usage or sales after initial registration. How is this driven by website design changes? Do night, workday, or weekend registrants have different lifetime value? Do new signup discounts hurt or help lifetime value?
- **HR**: Employee attrition. Are employees stickier during recessions? Do hires during certain periods or policies stay longer? How do company policy changes affect retention?
- **Sales staff**: Are new sales staff on track? How have product or pricing changes affected sales performance? Who are the true top performers adjusting for all this?
- **Movies**:
- **Store sales, etc. etc. etc.**: Just look for the vintage…

CONCLUSION

This article is meant to introduce Age-Period-Cohort models to those unfamiliar with them, but also to demonstrate the breadth of their possible application through a range of examples across different business and scientific applications. Age-Period-Cohort models are naturally suited to a range of statistical problems, including many that are essential for today's businesses. Not only are they intuitive in many contexts, they are a natural extension of the better-known survival models.

As seen in the SETI@home example, multiple APC models for different key rates can be used together in a system of equations to form a complete picture of lifetime value. APC models can be combined with other scoring or data mining techniques in order to answer critical account-level questions without loosing the long-term impacts of vintage effects.

REFERENCES

Briffa, K. R., P. D. Jones, F. H. Schweingruber, W. Karlén and S. G. Shiyatov**,** 1996, "Tree- Ring Variables as Proxy-Climate Indicators: Problems with Low-Frequency Signals", in P. D. Jones, R.S. Bradley, J. Jouzel (eds), *Climate Variations and Forcing Mechanisms of the Last 2000 Years* (New York: Springer), pp. 9–41.

Briffa, K. R., T. J. Osborn, F. H. Schweingruber, I. C. Harris, P. D. Jones, S. G. Shiyatov and E. A. Vaganov**,** 2001, "Low-Frequency Temperature Variations from a Northern Tree Ring Density Network", *Journal of Geophysical Research* 106, pp. 2,929–41.

Breeden, J.L. 2007. "Modeling Data with Multiple Time Dimensions", *J of Computational Statistics and Data Analysis*, Vol. 51, Issue 9, pp. 4761-4785.

Breeden, J.L., L. Thomas, and J.W. McDonald III. 2008. *"Stress-testing retail loan portfolios with dual-time dynamics," The Journal of Risk Model Validation,* ***2****(2),* Summer*, pp. 43-62.*

Breeden, J.L. 2010. "Testing retail lending models for missing cross-terms", *Journal of Risk Model Validation*, **4**(4), Winter.

Breeden, J.L., 2013. "Incorporating Lifecycle and Environment in loan-level forecasts and stress tests", *Proceedings of the Credit Scoring and Credit Control Conference XII*, *Edinburgh, 2013.*

Breeden, J.L. 2014. *Reinventing Retail Lending Analytics: Forecasting, Stress Testing, Capital, and Scoring for a World of Crises – Second Impression*, Riskbooks.

Breeden, J.L., A. Bellotti, and A. Yablonski. 2015. "Instabilities using Cox PH for forecasting or stress testing loan portfolios". *Proceedings of the Credit Scoring and Credit Control Conference XIV, Edinburgh, 2015.*

Breeden, J.L. and J. Canals-Cerdá. 2016. "Consumer risk appetite, the Credit Cycle, and the Housing Bubble", *Working Papers, Research Department, Federal Reserve Bank of Philadelphia*, No. 16-05.

Breeden, J.L. and L.C. Thomas. 2016. "Solutions to Specification Errors in Stress Testing Models", to appear in Journal of the Operational Research Society, 2016, doi:10.1057/jors.2015.97.

Cook, E. R., and L. A. Kairiukstis**,** 1990, *Methods of Dendrochronology* (Dordrecht: Kluwer Academic).

Holford, T.R. 1983. "The Estimation of Age, Period and Cohort Effects for Vital Rates", *Biometrics*, Vol. 39, No. 2 pp. 311-324

Grissino-Mayer, H., 1996, "A 2129-year reconstruction of precipitation for North-Western New Mexico, USA", in J. S. Dean, D. M. Meko and T. W. Swetnam (eds), *Tree Rings: Environment and Humanity* (Tucson, AZ: Radiocarbon), pp. 191–204.

Schmid V.J. and L. Held. 2007. *Journal of Statistical Software*, Volume 21, Issue 8. “Bayesian Age-Period-Cohort Modeling and Prediction – BAMP”.

Yang, Y. and K. C. Land. 2013. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications.* New York: Chapman & Hall / CRC.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Joseph L. Breeden
Prescient Models LLC
breeden@prescientmodels.com
www.prescientmodels.com