SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

# Forecasting Behavior with Age-Period-Cohort Models

## How APC Predicted the US Mortgage Crisis, but Also Does So Much More - 2700

# About the presenter

Dr. Breeden has been designing and deploying forecasting systems since 1994. He co-founded Strategic Analytics in 1999, where he led the design of advanced analytic solutions including the invention of Dual-time Dynamics.  He currently runs Prescient Models, which focuses on portfolio and account-level forecasting solutions for lifetime value assessment, account management, and stress testing.

Dr. Breeden has created models through the 1995 Mexican Peso Crisis, the 1997 Asian Economic Crisis, the 2001 Global Recession, the 2003 Hong Kong SARS Recession, and the 2007-2009 US Mortgage Crisis and Global Financial Crisis. These crises have provided Dr. Breeden with a rare perspective on crisis management and the analytics needs of executives for strategic decision-making.

He has published over 40 academic articles and 6 patents. The second edition of his book “Reinventing Retail Lending Analytics: Forecasting, Stress Testing, Capital, and Scoring for a World of Crises” was published by Riskbooks in 2014.

Dr. Breeden received separate BS degrees in mathematics and physics in 1987 from Indiana University. He earned a Ph.D. in physics in 1991 from the University of Illinois studying real-world applications of chaos theory and genetic algorithms.

# Talk Outline

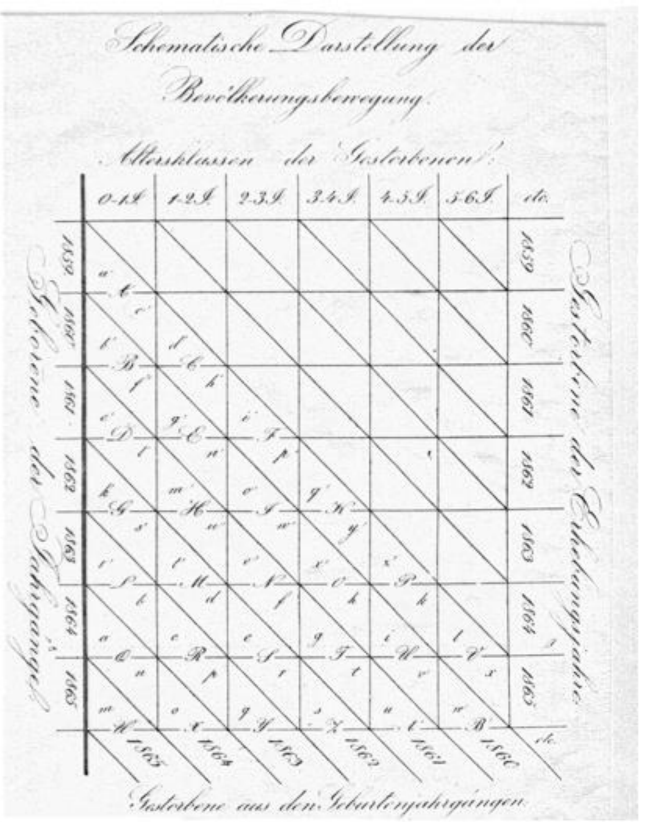
- What are Age-Period-Cohort models?
- What's the hidden story of the US Mortgage Crisis?
- Where else do APC models apply?
- Are there implementation details we need to know?

Definitions

# What are Age-Period-Cohort models?

# Lexis Diagrams

- Between 1869 and 1875, Zuener, Brasche, Becker and Lexis invented a way to look at mortality rates by separating the data into year of birth (cohort) and year of death (period). Age is of course the age at death (period – cohort).
- In the 1960s and 70s, a rich literature developed in demography and epidemiology on how to estimate functions of age, period, and cohort using aggregate data like in the Lexis diagram.



# Age Period Cohort (APC) Models

- Given an origination date (vintage), the age of the account is calendar date – vintage date, $a = t - v$.
- Functions of age $F(a)$, vintage $G(v)$, and time $H(t)$ are then estimated. For binomially distributed events, this is

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

- The functions are most commonly estimated parametrically via splines or nonparametrically. Generalized Linear Model (GLM) estimation for splines, and Bayesian, Partial Least Squares, or Ridge Regression are used for the nonparametric functions.

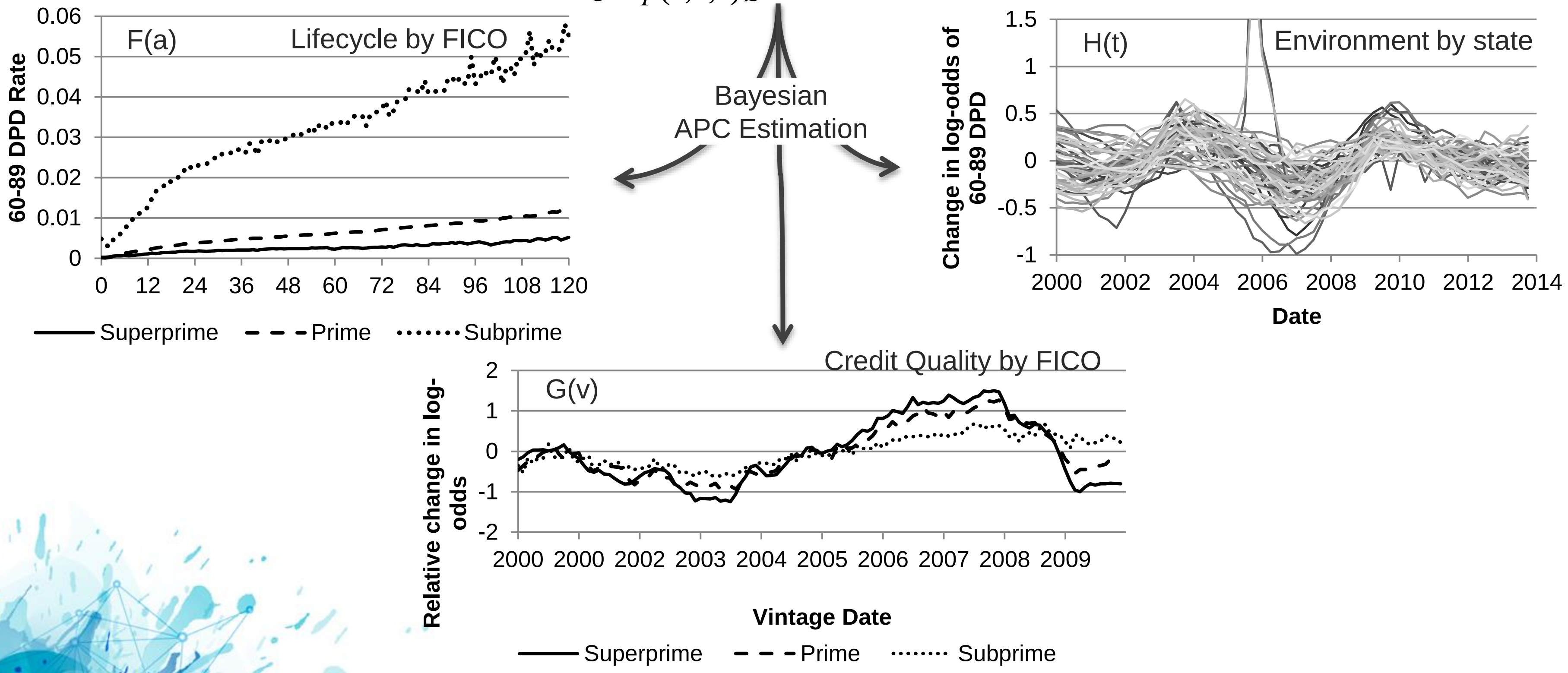
Use Case

# What's the hidden story of the US Mortgage Crisis?

Joint research with José Canals-Cerdá, Federal Reserve Bank of Philadelphia, jose.canals-cerda@phil.frb.org.

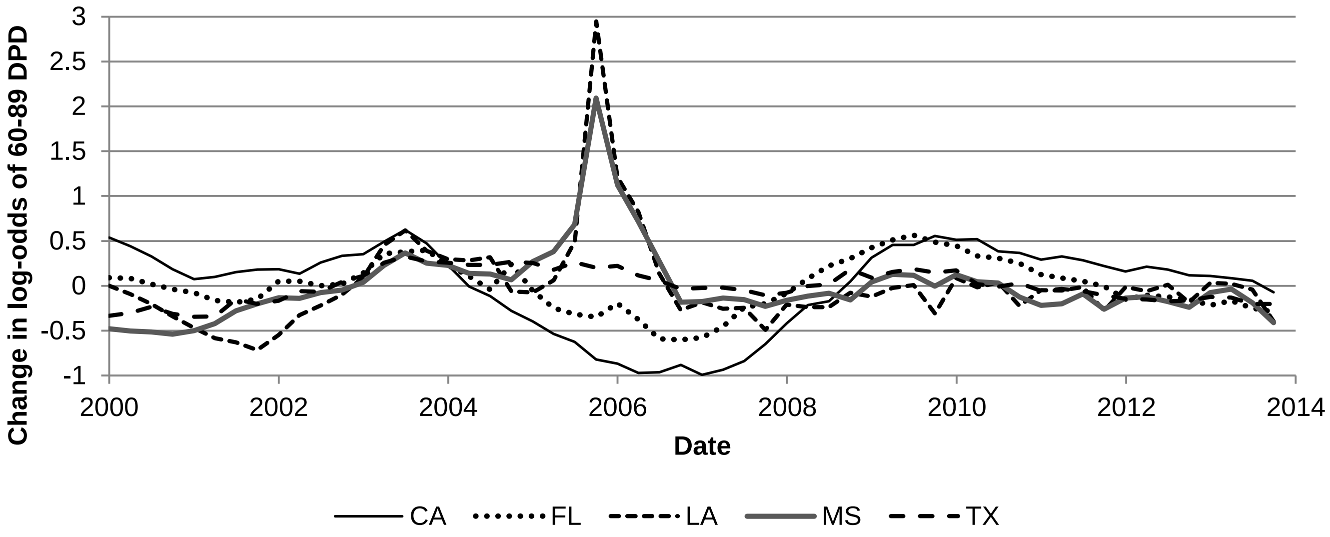
# Age-Period-Cohort Decomposition

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



# Adjusting for the Environment

- The environment function can reveal impacts for which no predictive factors are available.
- The example below shows Hurricane Katrina’s impact on mortgage delinquency. Later results are normalized any environmental impacts, like those shown here.



# Loan-level Modeling

- We use an APC decomposition initial step so that we capture all of the lifecycle and environment variation, and so that we control the linear trend ambiguity in age, vintage, time models. (More on this later.)
- Then we keep the lifecycle and environment as fixed offsets in a GLM score using quarterly performance data.
- We include typical origination scoring factors and then test for the inclusion of vintage fixed effects (dummy variables) $g_v$.

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

# Scoring Factors

The origination scoring factors and coefficients were typical for 1st lien, fixed rate mortgages.

We excluded all of the exotic products, e.g. Option-ARMs. We're analyzing traditional mortgage products.

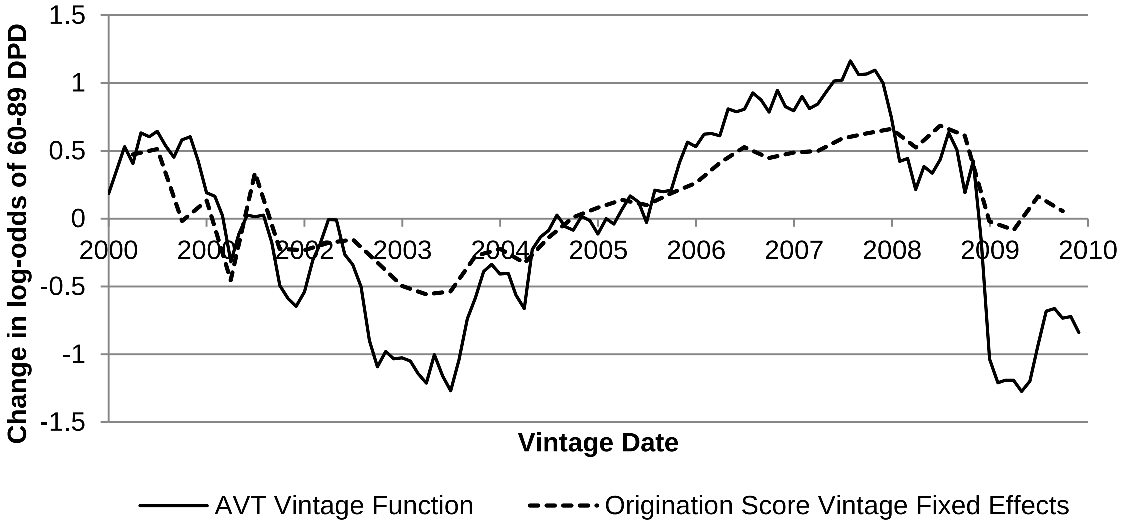
**Table 4: Output Coefficients from the GLM Analysis of Mortgage Delinquency.**

| variables | Coef. | t-val | Variables (cont.) | Coef. | t-val |
|---|---|---|---|---|---|
| Intercept | 2.268 | 67.83 | **Source Channel** | | |
| Jumbo Loan | -0.128 | -16.57 | source 1 | 0 | |
| **Documentation** | | | source 2 | 0.217 | 63.27 |
| Full documentation | 0 | | source 7 | 0.100 | 30.25 |
| Low documentation | 0.103 | 25.48 | source T | 0.240 | 38.63 |
| No documentation | -0.030 | -5.16 | source U | 0.437 | 32.19 |
| Documentation unknown | 0.135 | 40.61 | **Occupancy** | | |
| **Fico at Origination** | | | Owner | 0 | |
| up to 540 | 0 | | Non-owner | -0.109 | -17.75 |
| 540 to 580 | -0.188 | -17.33 | Other or Unknown | -0.187 | -17.84 |
| 580 to 620 | -0.435 | -44.66 | **PMI** | | |
| 620 to 660 | -0.807 | -85.86 | No | 0 | |
| 660 to 700 | -1.373 | -145.22 | Yes | 0.119 | 32.06 |
| 700 to 740 | -1.956 | -202.66 | Unknown | 0.170 | 37.92 |
| 740 to 780 | -2.671 | -264.46 | **Term** | | |
| 780 to 820 | -3.380 | -283.96 | 0 to 120 | 0 | |
| 820+ | -3.623 | -52.71 | 120 to 180 | 0.144 | 10.39 |
| **Loan to Value** | | | 180 to 240 | 0.407 | 27.38 |
| 0 to 0.75 | 0 | | 240 to 360 | 0.593 | 44.05 |
| 0.75 to 0.8 | 0.157 | 40.09 | 360+ | 0.640 | 36.25 |
| 0.8 to 0.85 | 0.221 | 46.80 | **Purpose** | | |
| 0.85 to 0.9 | 0.247 | 42.01 | Purchase | 0 | |
| 0.9 to 0.95 | 0.262 | 42.53 | Refinance | -0.001 | -0.41 |
| 0.95 to 1 | 0.305 | 46.79 | purpose U | -0.462 | -64.29 |
| 1 to 1.13 | 0.285 | 36.81 | purpose Z | 0.090 | 5.08 |
| DTI | 0.007 | 73.35 | | | |

**Note: The model specification includes also quarterly vintage dummies that are not explicitly reported in this table.**

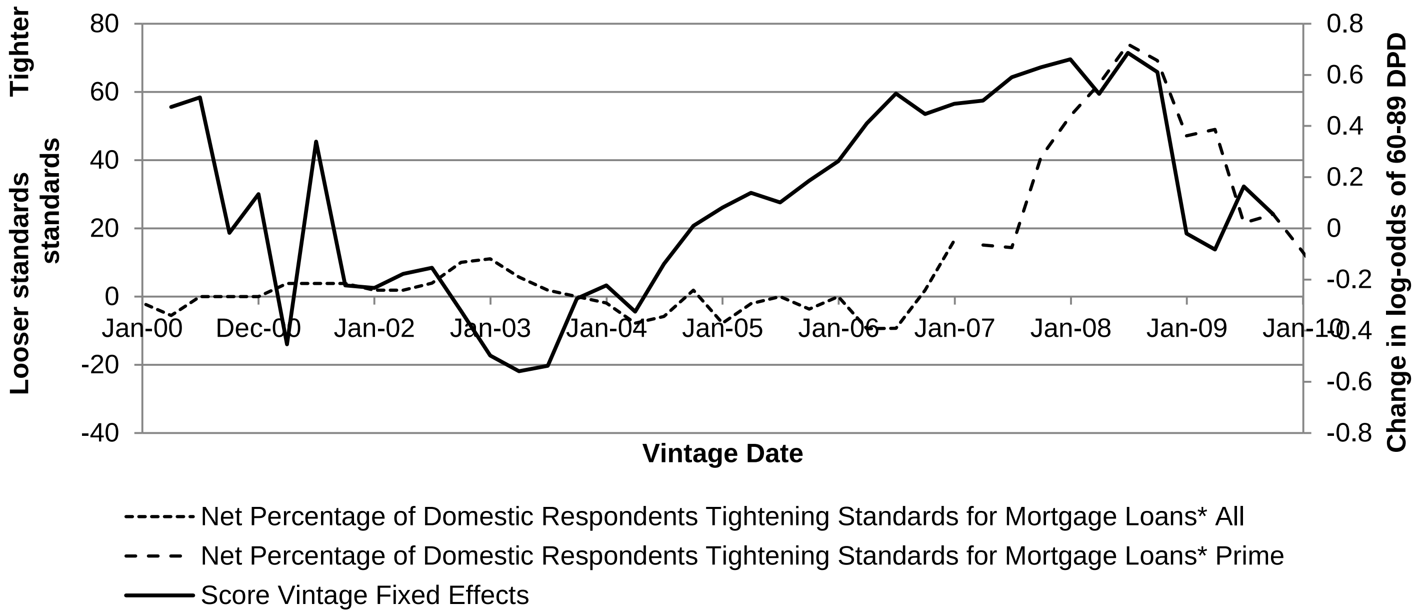
# APC Vintage Function versus Scoring Vintage Dummies

- The APC vintage function measures the net impact to log-odds of default for the variation in credit risk by vintage.
- The scoring vintage dummies measure the vintage residual after scoring factors.
- By comparing, we see that only half of the vintage variation is explainable by scoring factors.



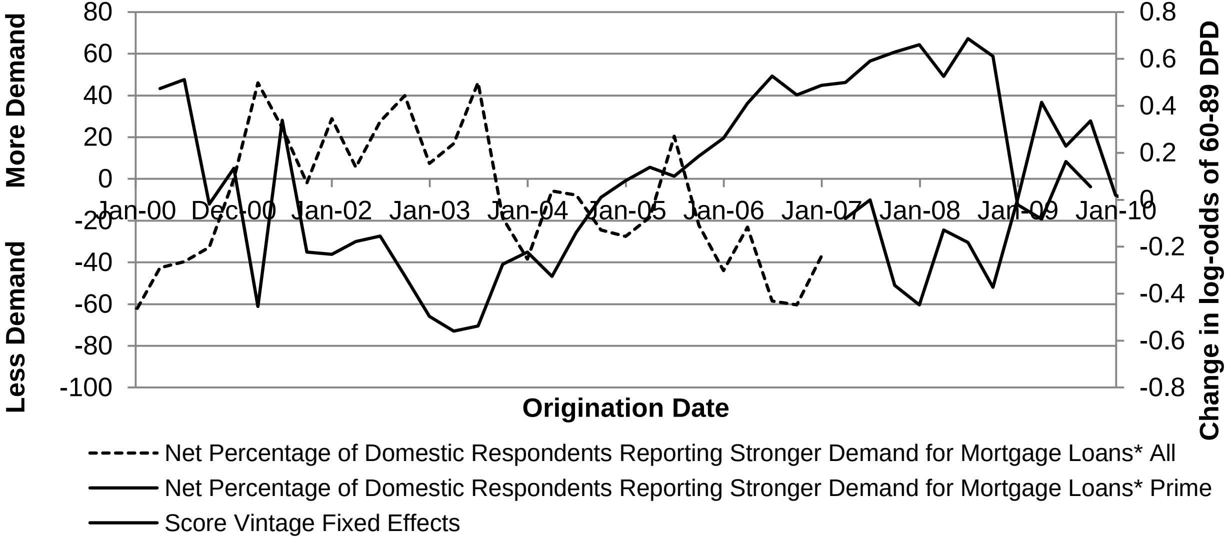
# FRB SLOOS Survey – Underwriting Standards

- The Federal Reserve publishes a Senior Loan Officer Opinion Survey (SLOOS).
- Self-reported changes in underwriting standards show a correlation of $\rho$ =0.4 ± 0.4 to the vintage fixed effects, with the wrong sign.



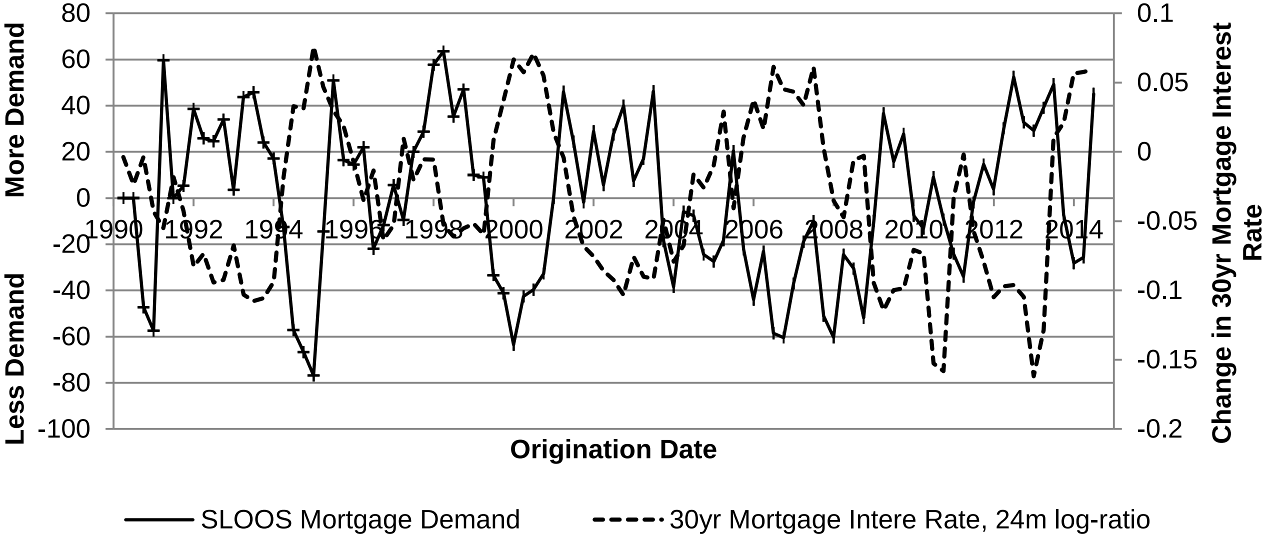
# FRB SLOOS Survey – Consumer Loan Demand

- The same senior loan officers in the same survey report on consumer demand for loans.
- Consumer demand has a correlation of $\rho = -0.7 \pm 0.3$ to the vintage effects.
- When consumer demand is high, the loans are good – consumer risk appetite is a real driver of credit risk.



# Drivers of Consumer Demand

- Historically, SLOOS-reported consumer mortgage demand is highly correlated to the 24-month change in mortgage interest rates.
- When offered interests fall over a sustained period, consumer demand rises.



Applications

# Where else do APC models apply?

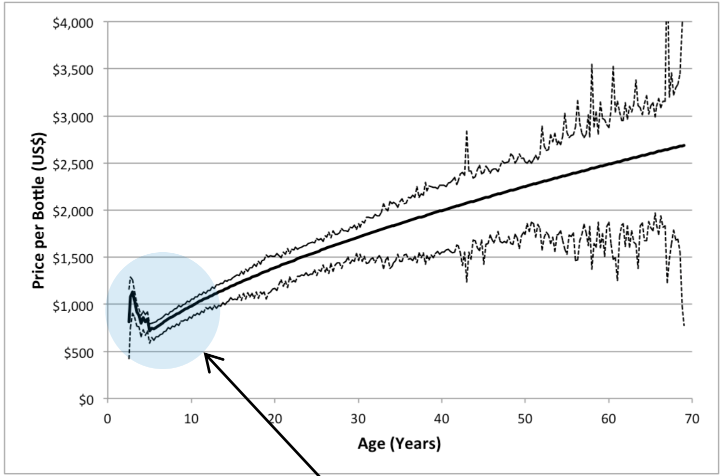
# Other Applications of APC Models

- Wine Forecasting
- SETI@home
- Dendrochronology
- eCommerce customer lifetime value
- HR
- Sales

# Fine Wine Value Forecasting

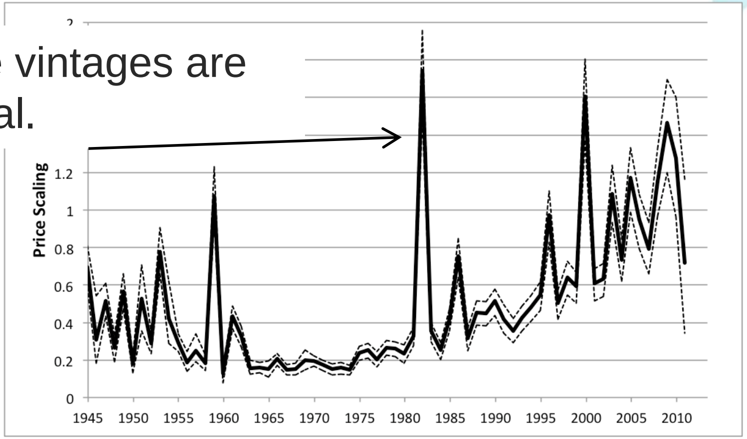
- Auctionforecast.com – Vincast
- 1.5 million auction prices over a 10 year period
- Predicts wine price, starting with a Bayesian APC decomposition.
- Wine is perfect for “vintage” analysis.

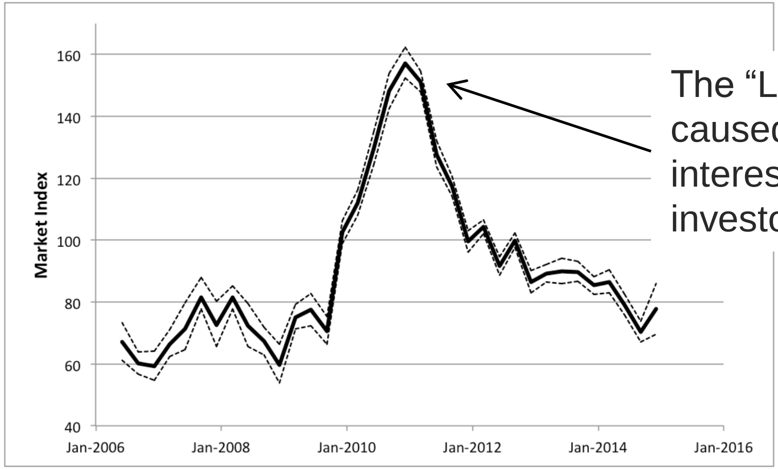
# Vincast Decomposition – Chateau Lafite Rothschild




Prices actually drop through the first 5 years from release.

Some vintages are special.

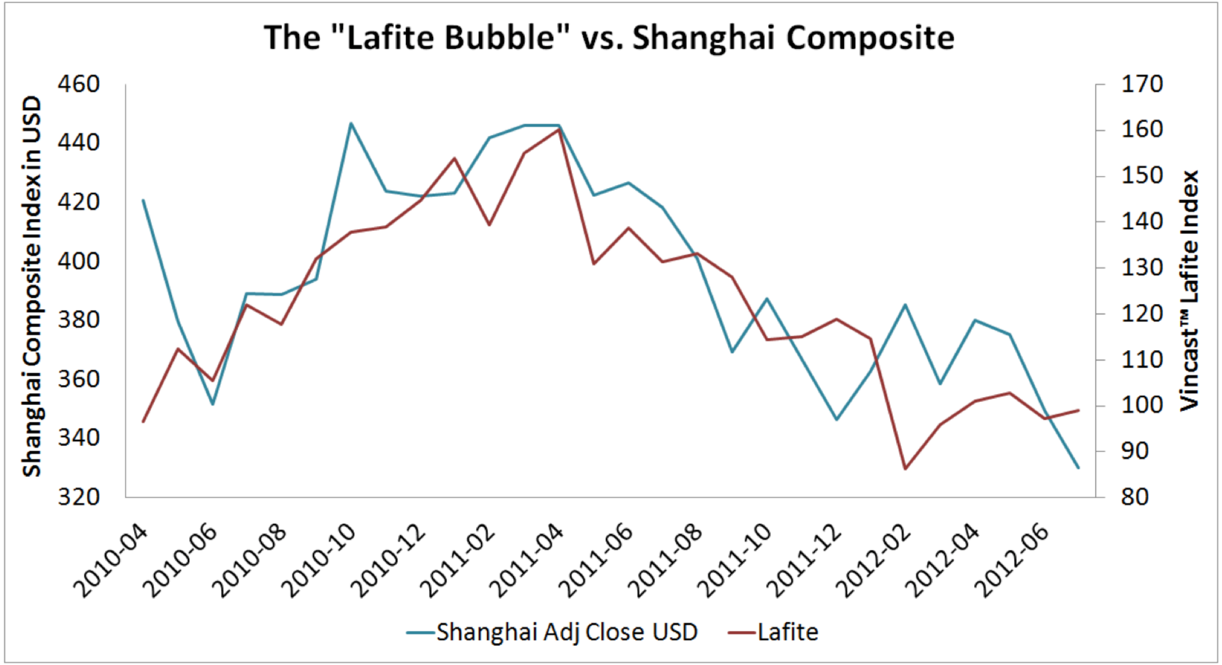



The “Lafite Bubble”, caused by a flurry of interest from Chinese investors.

# Drivers of the Wine Market

- The environment function (market index) for auctions prices has tracked Chinese wealth for the last decade.

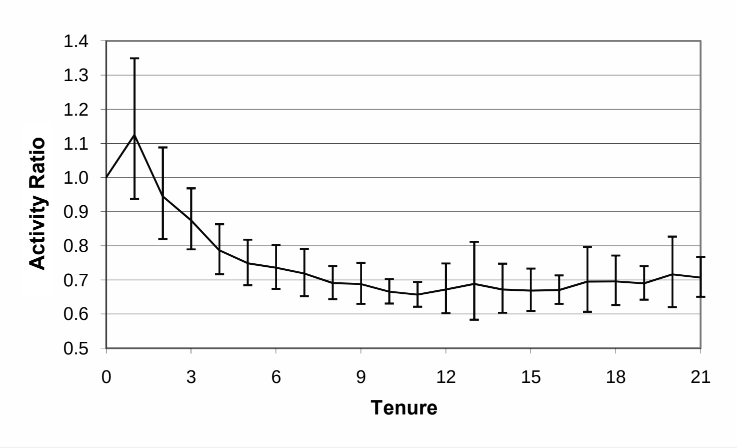


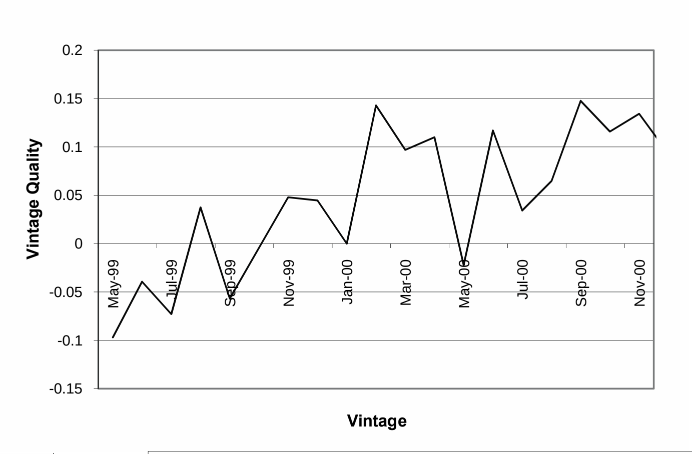
# SETI@home Member Analysis

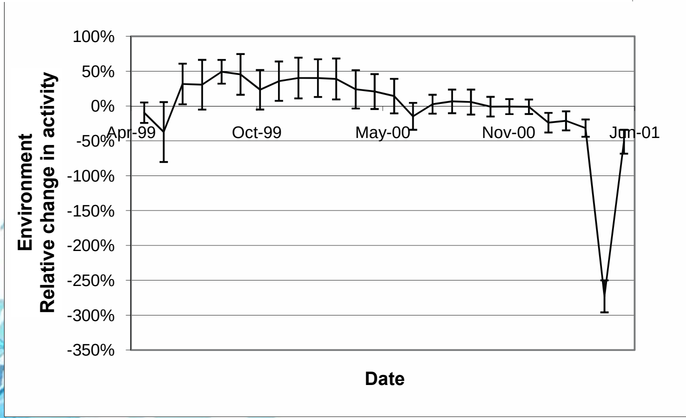
- SETI@home is the Search for Extraterrestrial Intelligence signal detection project.
- We analyzed 2 years of member data, analyzing activity rates, number of cpus applied, and analysis return rates. Leading to member lifetime value estimates segmented by Country, OS, and software version.
- As an example of the lifetime value analysis, we show load, equivalent to usage for a website or spend on a credit card.

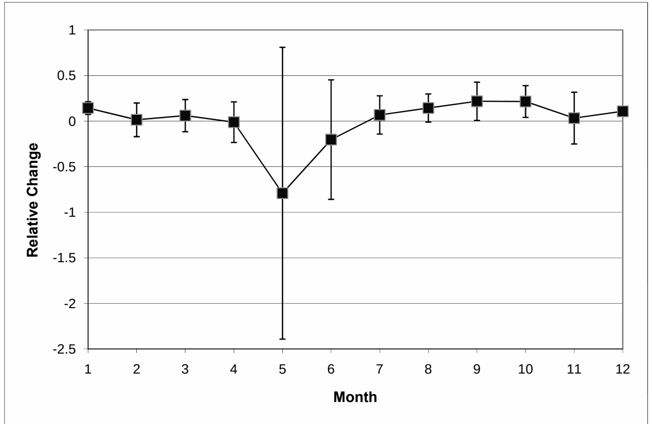
# Activity Decomposition

$$activity(a,v,t) = \frac{active\,accounts(a,v,t)}{active\,accounts(a=0,v,t)}$$



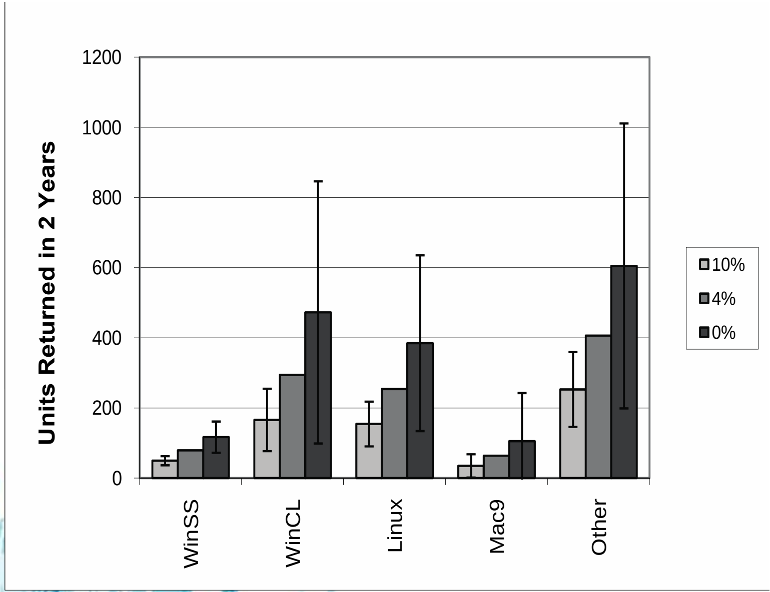



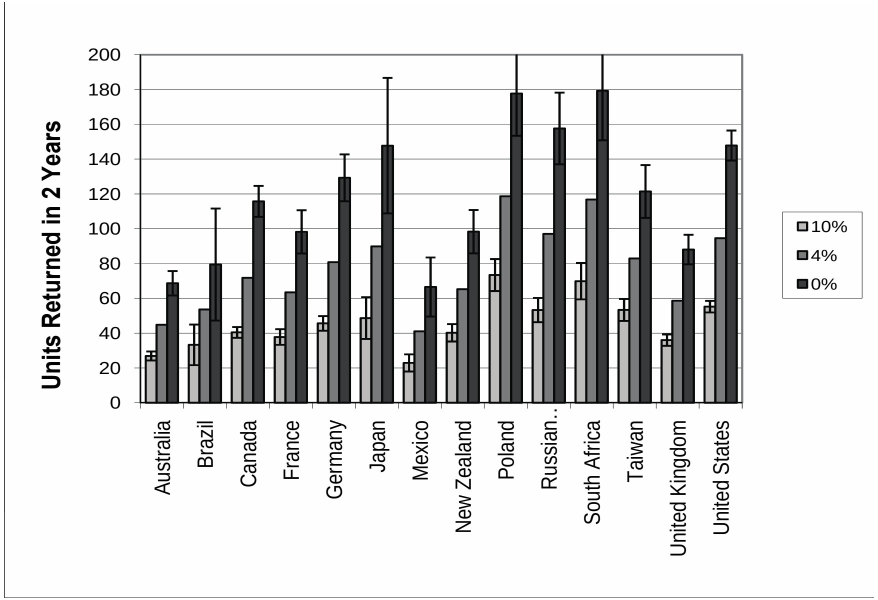



# Member Net Present Value

- Combining all factors, we predicted NPV for new members, allowing the SETI@home managers to target their development by OS and country.



# Dendrochronology – Tree Ring Analysis

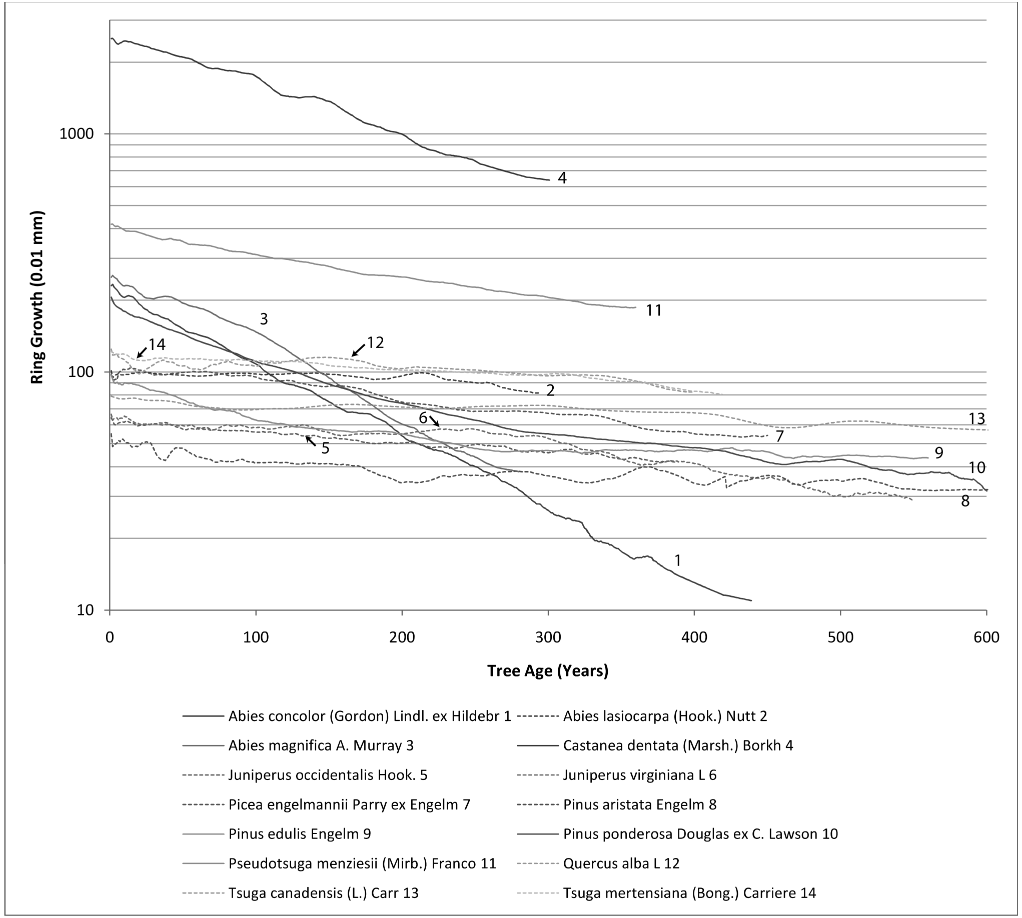
- Tree ring growth rates are driven by age of the tree, environmental conditions, and growth conditions for the individual tree.

| **Species** | **Number of Trees** |
|---|---|
| Abies concolor (Gordon) Lindl. ex Hildebr | 302 |
| Abies lasiocarpa (Hook.) Nutt | 3652 |
| Abies magnifica A. Murray | 241 |
| Castanea dentata (Marsh.) Borkh | 177 |
| Juniperus occidentalis Hook. | 1164 |
| Juniperus virginiana L | 630 |
| Picea engelmannii Parry ex Engelm | 1562 |
| Pinus aristata Engelm | 256 |
| Pinus edulis Engelm | 2075 |
| Pinus ponderosa Douglas ex C. Lawson | 4998 |
| Pseudotsuga menziesii (Mirb.) Franco Quercus alba L | 5324 |
| Quercus alba L | 2690 |
| Tsuga canadensis (L.) Carr | 1538 |
| Tsuga mertensiana (Bong.) Carriere | 1027 |

# Tree Ring Decomposition – Lifecycles

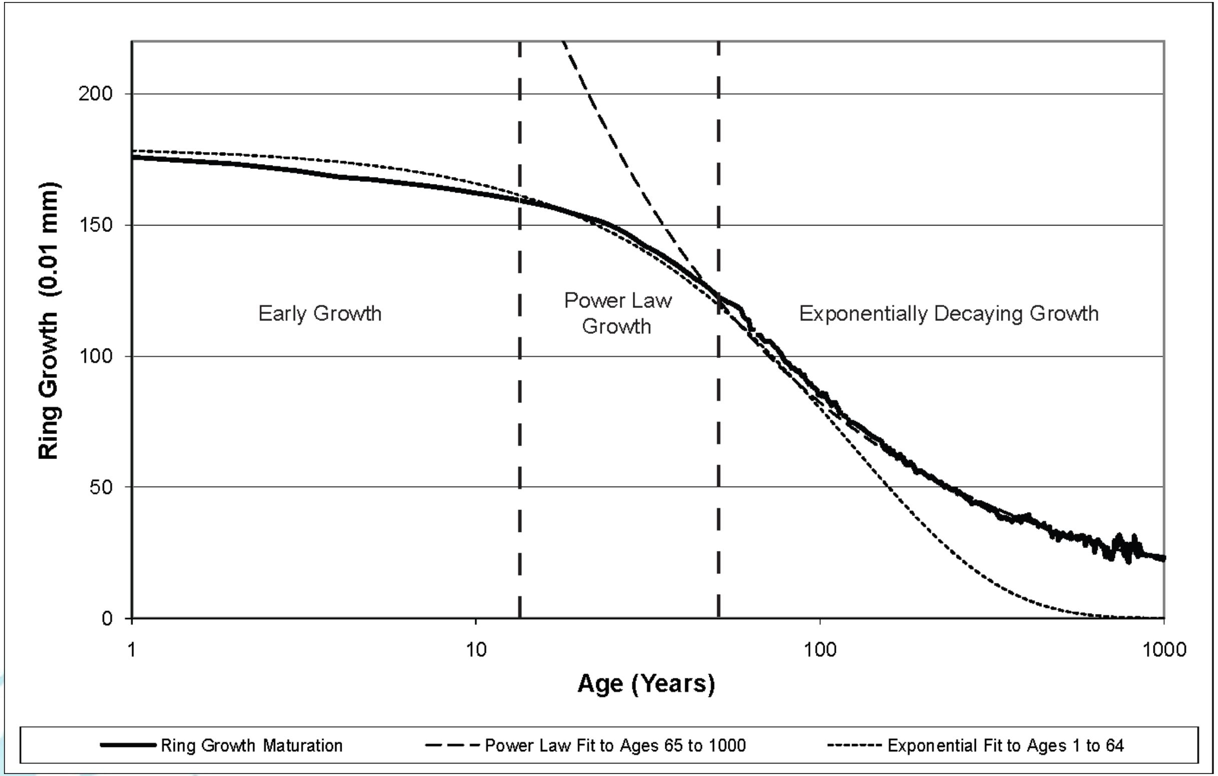
$$\log(w) = F(a) + G(v) + H(t)$$

Growth rate patterns differ by species.



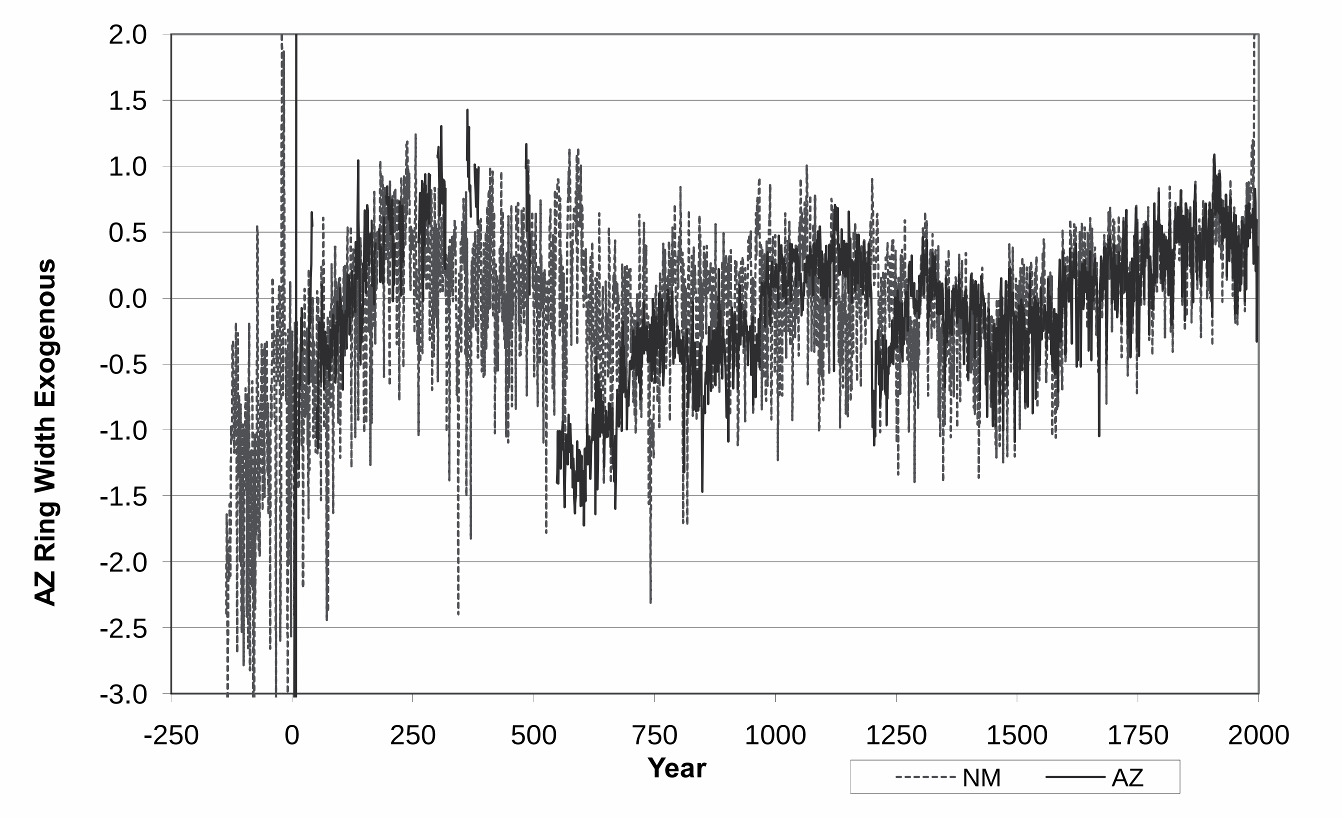
# Tree Ring Growth Lifecycle Phases

- Tree ring growth goes through different phases. Nonparametric estimation allows us to discover this.



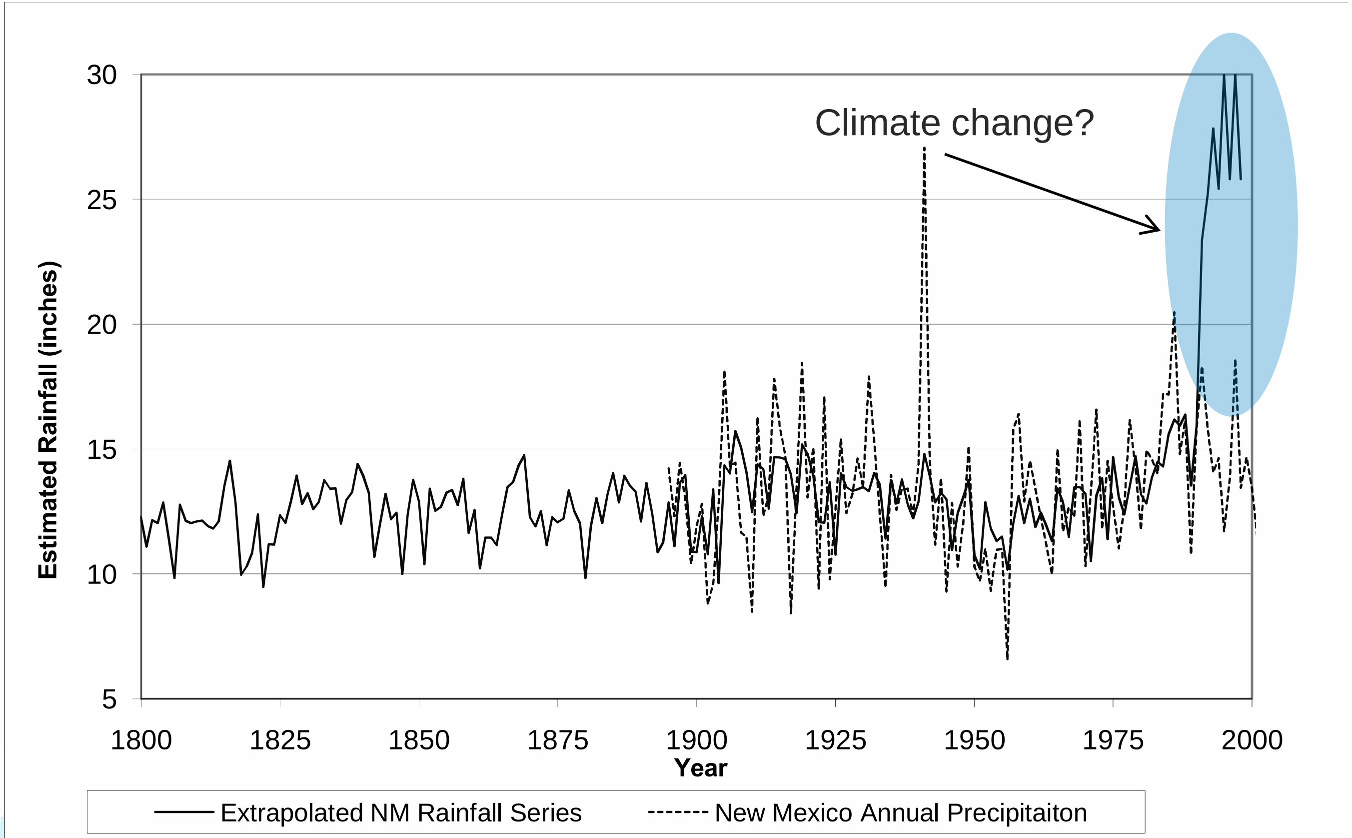
# Tree Ring Decomposition – Environment

- The environment functions for New Mexico and Arizona show similar patterns through the centuries.



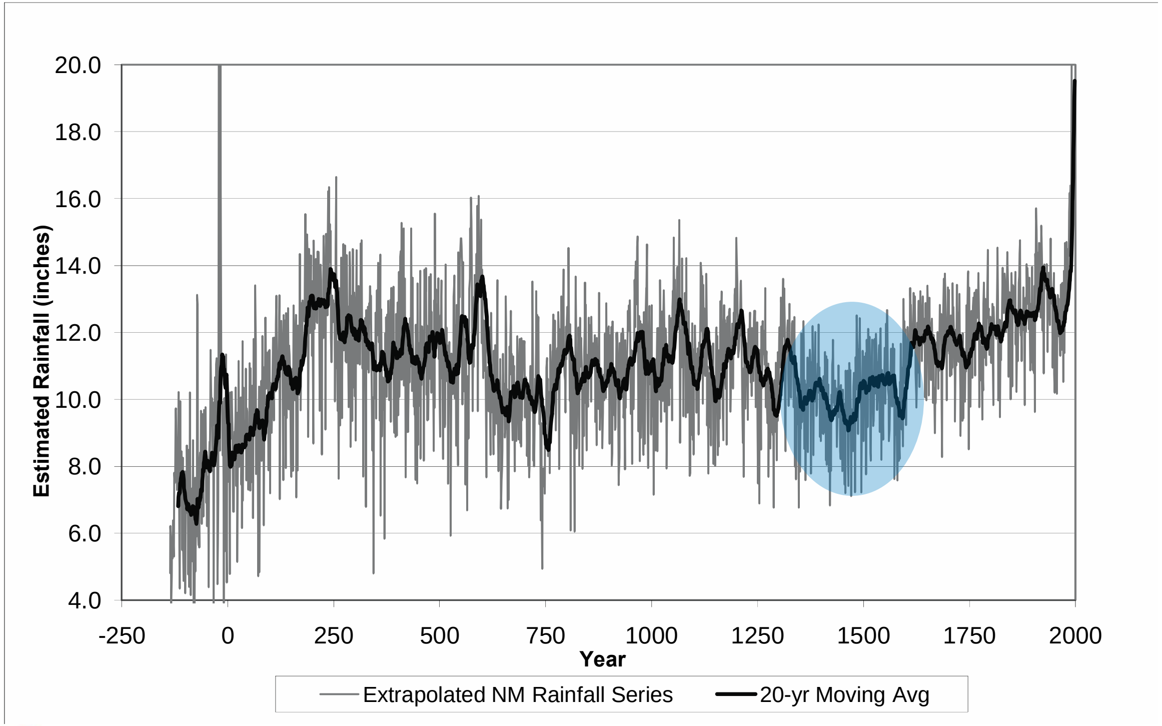
# Environmental Correlations

- Recent ring growth correlates well to rainfall, until the 1980s.



# Predicting Historic Rainfall Levels

- Backward-extrapolating the rainfall correlation shows prehistoric droughts, possibly explaining the fall of the Anasazi civilization.



# eCommerce, HR, Sales, etc.

- Any data set where a vintage can be defined and performance tracked with time can be analyzed via APC
  - **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?
  - **Store sales, etc. etc. etc.**: Just look for the vintage…

Implementation

# Are there implementation details we need to know?

# Implementation Details

- **Link Functions**: Choose the correct distribution to match your data. Estimators are readily available for binomial, Gaussian, and lognormal.
- **Estimation technique**:
  - Spline estimation is available via `proc transreg`.
  - Bayesian estimation 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).`
- **Cross-terms**: Test the decomposition to see if the age, vintage, and time functions are independent. If they are not, try segmentation.

# Linear Trend Ambiguity

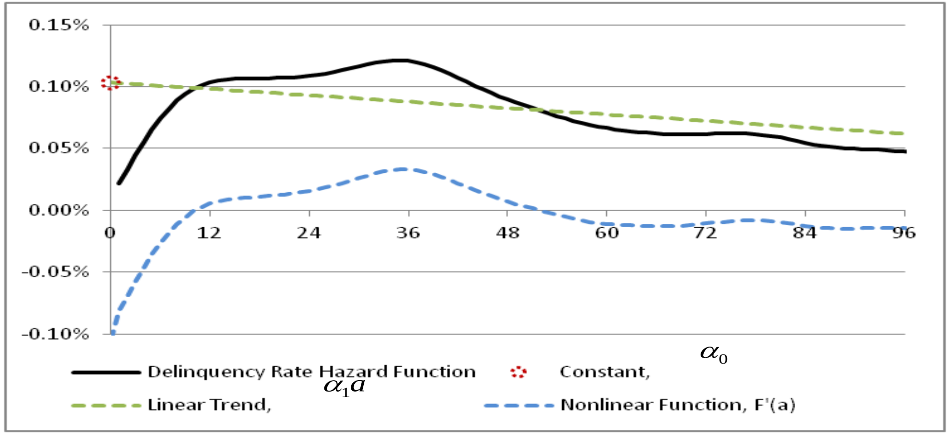
- Because a=t-v, there is an ambiguity in the allocation of linear trends. Consider

$$\ln r(a,v,t) = F(a) + G(v) + H(t) + e(a,v,t)$$

- These functions are measured on a discrete number of months. Therefore, they can be represented precisely with a polynomial through those points.
- We can separate the constant, linear, and nonlinear terms as

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

Constant Linear Nonlinear



# The Constant Terms

- Substituting the polynomial forms into the original expression, we get

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

- This shows that we have three constant terms, where only one can be estimated uniquely. This can be remedied by the following definition:

$$\alpha_0' = \alpha_0 + \beta_0 + \gamma_0$$
$$\beta_0' = 0$$
$$\gamma_0' = 0$$

- Rewriting gives:

$$\ln r(a,v,t) = \alpha_0' + \alpha_1 a + F'(a) + \beta_1 v + G'(v) + \gamma_1 t + H'(t) + \varepsilon(a,v,t)$$

# The Linear Terms

- Because of the relationship between age, vintage, and time: $a = t - v$

  or equivalently $t = v + a$

- We can write

$$\ln r(a,v,t) = \alpha_0' + \alpha_1 a + F'(a) + \beta_1 v + G'(v) + \gamma_1 (v + a) + H'(t) + \varepsilon(a,v,t)$$

- With the definition

$$\ln r(a,v,t) = \alpha_0' + (\alpha_1 + \gamma_1) a + F'(a) + (\beta_1 + \gamma_1) v + G'(v) + H'(t) + \varepsilon(a,v,t)$$

  this becomes $\alpha_1' = \alpha_1 + \gamma_1, \beta_1' = \beta_1 + \gamma_1, \gamma_1' = 0$

$$\ln r(v,a,t) = a_0' + a_1' a + F'(a) + b_1' v + G'(v) + H'(t) + e(a,v,t)$$

# Linear Interpretation

- Because of the relationship between age, vintage, and time, we cannot have three unique linear terms, only two.
- Known from the Age Period Cohort literature, it means that we can never be certain of the magnitude of the linear trends in lifecycle, environment, or credit risk.
- This result is true for all analysis of vintage data, regardless of the model used.
- This problem is solved by making domain-specific assumptions.

# Conclusions

- Age-Period-Cohort models are naturally suited to a range of statistical problems, including many that are essential for today’s businesses.
- APC models can be used together to create a complete picture of net present 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


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

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Contact info:

breeden@prescientmodels.com

+1-505-670-7670