THE IMPACT OF BIG DATA ON FIRM PERFORMANCE: AN EMPIRICAL INVESTIGATION

PATRICK BAJARI*, VICTOR CHERNOZHUKOV†, ALI HORTAÇSU⋆, AND JUNICHI SUZUKI‡

ABSTRACT. In academic and policy circles, there has been considerable interest in the impact of "big data" on firm performance. We examine the question of how the amount of data impacts the accuracy of Machine Learned models of weekly retail product forecasts using a proprietary data set obtained from Amazon. We examine the accuracy of forecasts in two relevant dimensions: the number of products (N), and the number of time periods for which a product is available for sale (T). Theory suggests diminishing returns to larger N and T , with relative forecast errors diminishing at rate $1/\sqrt{N}+1/\sqrt{T}$. Empirical results indicate gains in forecast improvement in the T dimension; as more and more data is available for a particular product, demand forecasts for that product improve over time, though with diminishing returns to scale. In contrast, we find an essentially flat N effect across the various lines of merchandise: with a few exceptions, expansion in the number of retail products within a category does not appear associated with increases in forecast performance. We do find that the firm's overall forecast performance, controlling for N and T effects across product lines, has improved over time, suggesting gradual improvements in forecasting from the introduction of new models and improved technology.

Date: February 14, 2018.

* University of Washington, † MIT, ⋆ University of Chicago, ‡ Amazon, Inc.

Bajari is VP and Chief Economist at Amazon, Inc. Chernozhukov and Hortaçsu carried out research work for this paper as independent contractors for Amazon, Inc. We thank Zhiying Gu and Liquan Huang with assistance in data collection and analysis, and seminar participants at INFORMS, Bruegel, The World Trade Organization and UCLA for insightful comments.

1. Introduction

The use of data as an input into the production function of a firm dates back at least to the emergence of the modern industrial firm in the 19th century (Chandler (1977)). Indeed, the first mention of the word "business intelligence" is in Devens' Cyclopaedia of Commercial and Business Anecdotes' in 1865. The dramatic drop in the cost of computation technologies since the 1990s has made it much easier for firms to collect, store, and analyze data to help with their decision processes, leading to changes in organizational and management practices, and hence productivity growth (see e.g. Bresnahan, Brynjolfsson, Hitt (1994)), Tambe and Hitt (2012), Bloom et al. (2012), McElheran and Jin (2017)).

In this paper we study the performance of Amazon's retail forecasting system, an important use of data within the firm. At any point in time, there are a large number of unique retail products in Amazon's warehouses supplied by vendors. Each product has a forecast produced internally. The forecast is a probability distribution function for the current week and up to 52 weeks into the future.

Amazon's automated purchase ordering systems use forecasts as an important input. Forecast accuracy is important for the performance of Amazon's retail business. If a forecast is downward biased (i.e. the forecast is less than the actuals), Amazon loses short run sales and profits from going out of stock. Also, it risks disappointing customers and they may be less likely to make future purchases. If the forecasts are upward biased (i.e. the forecast is greater than the actuals), Amazon will have inefficient inventory turns and may need to markdown or liquidate overstock products.

Our analysis will build on a proprietary panel data set on 5 years of weekly historical forecasts and actual demands for 36 major product lines including apparel, books and consumer electronics. Our ability to observe the Amazon's internal demand forecasts along with data on actual product sales allows us to define a very clear measure of process improvement: does having access to more data allow Amazon to reduce forecast error?

This question has also recently attracted attention in some policy discussions, with the business press and research literature advancing the "data feedback loop" hypothesis. This can be interpreted as the presence of an indirect network effect, where the accumulation of bigger data sets by firms, through the addition of more users and/or products, helps to improve their products and services. This, in turn, attracts more users/products, and thus access to more data. In the context of forecasting, one might interpret this as with more data, firms can produce better forecasts, which in turn allows them to better serve customers, which in turn leads to more data. There are a number of theoretical discussions of the "data feedback loop" hypothesis e.g. Newman (2014), Grunes and Stucke (2015), Lerner (2014), and Lambrecht and Tucker (2017). However, to the best of our knowledge, there is no empirical research that examines the performance of a large scale software system within the modern corporation to determine how adding more data changes forecast errors.

We consider four hypotheses of factors which might influence the accuracy of forecasts. First, statistical learning theory typically suggests diminishing returns to data set size in terms of estimation error or predictive performance (Lerner (2014)). In our particular setting, the underlying data is a panel, and the theory suggests that forecast accuracy should improve as we change N, the number of products within a product category. Second, we would expect forecast errors to decrease as T, the number of observations per product, increases. Third, as pointed out by Bresnahan, Brynjolffson, and Hitt (2004), Bloom et al. (2012), Lambrecht and Tucker (2017), complementary investments and organizational practices often play a very important role in generating productivity gains from the use of data. In our particular setting, there is learning by doing on the forecasting team as they build new models, use improved hardware due to improvements in Amazon Web Services, and concurrently improve their organizational practices. Finally, there can be "diseconomies of scale" in forecasts. As the amount of data grows, the team

may need to apply the same model to an increasing number of products with different demand patterns. If Amazon lacks the headcount to customize the forecasting models to a specific group of products, it will need to use a more "one-size fits all" modeling approach and accuracy may suffer.

The first step in our analysis is to develop a theory of forecast errors for our application. In practice multiple models are used and they are constantly changing. Rather than characterize the behavior for a specific production model, our theoretical benchmark model is based on a state of the art forecasting paradigm applicable to panel data sets such as ours. In particular, in Section 2, we use the interactive fixed effect model, also known as the augmented factor model, following, e.g., Bernanke et al (2004) and Bai (2009). We characterize the asymptotic distribution of forecast errors and show that they shrink to the irreducible level at the rate $1/\sqrt{N} + 1/\sqrt{T}$. We note that in practice, the underlying data is high dimensional and the teams need to engage in model and variable selection. Asymptotic theory suggests that the rate of convergence may be slower as a result.

We next use panel data models to estimate forecast errors as a function of N, T and other factors. Our dependent variables are measures of forecast accuracy (e.g. making a percentage forecast error greater than some threshold). Asymptotic theory suggests the rate at which forecast errors decrease towards the irreducible level, and we use this to control for N and T parametrically. We use time effects to estimate flexible models that control for trend and seasonality. If there is learning by doing in forecasting, we would expect to see a negative time effect.

Our results suggest that there is robust improvement in forecast performance in the T dimension (though subject to diminishing returns, agreeing with the theoretical prediction). This is perhaps unsurprising given that many product lines have highly unbalanced panels with half or more of the products entering and exiting in a given year. For a new product, T may be near zero and even a small number of observations may be impactful in terms of reducing forecast errors.

Our results suggest that N has relatively little impact on forecasting performance, except when the number of products is relatively small (e.g. a completely new product line with only a few thousands unique products). This is largely consistent with what we expect from theory, which tells us that prediction errors will typically exhibit diminishing returns in the relevant dimensions of data.

Finally, our results suggest that there is a trend rate of improvement in our forecasts. This is consistent with "learning by doing" where the team improves their models through a process of trial and error. Also, improvements in infrastructure enable them to train increasingly complicated models and more efficiently scale their forecasting systems.

We note that our results are inconsistent with a naive "data feedback loop," where the addition of new products always results in better models. We interpret our results as instead consistent with the statistical theory for the size of the forecasting errors. Our data is relatively scarce in the T dimension, particularly for new products. Increasing T from 1 to 52 weeks should be expected to meaningfully improve accuracy. However, in the N dimension, our empirical results show that having "Big Data" across a large array of products has little value in improving model accuracy. This could be because N has diminishing effects as suggested by the theory, and, as a consequence, scientists often do not use all of the data to train their models (focusing on random subsamples instead). For example, Varian (2014) suggests that Google uses only 0.1% subsamples of its data to power its decision-support systems. In our discussions with the forecasting team, we also learned that the team does not use all of the data to train their models, and that the accumulation of "Big Data" across a growing variety of products is often viewed as an engineering challenge rather than a modeling benefit. The teams need to worry about scaling challenges to run and vend the models to downstream systems. As the amount of data increases, the engineering challenges will become more difficult and more advanced solutions need to be deployed. In addition, learning from additional products are often limited as many of these new products do not sell at all for a period of time.

There are important limitations of our study. As we noted above, the T effect that we identify may not be the true causal effect of having access to longer histories. This could be driven also by the continual improvements in forecasting technology over time. As we discuss in Section 3.2, what we can recover, however, are bounds on the true effect of T.

Another important caveat we should emphasize is regarding the scope of this study. Clearly, data is used in many aspects of company decision making, and our focus is on the application to demand forecasting. We believe that this is a particularly good area to study, as the success of forecasting models are relatively straightforward to assess. However, our conclusions regarding the presence of scale benefits to data are limited to this particular application.

As we noted, there are not, to our knowledge, a large number of empirical studies testing the "scale effect of data" hypothesis. Lerner (2014) is a theoretical piece that suggests the inverse square root relationship, as arising from asymptotic theory. As noted above, Varian (2014) suggests that scale effects may not be important, as Google uses only 0.1% subsamples of its data to power its decision-support systems. Lambrecht and Tucker (2017) point out that often the algorithm used rather than the size of the data set is what improves prediction performance. De Fortuny et al. (2013) discuss the performance of a variant of Naive Bayesian classification algorithm on a number of data sets from online content providers as well as a bank. They find that the performance of the classifier algorithm continues to improve, with some evidence for diminishing returns, even after the number of data entries exceeds, in some cases, 500 million.

The rest of the paper proceeds as follows: in Section 3 we lay out a theoretical model specifying the data generation process for the stochastic demand faced by Amazon. We also specify the benchmark forecasting model and investigate its asymptotic properties. In Section 3, we discuss the data. Section 4 reports detailed results from the Electronics category. Section 5 reports results from all 36 product categories. Section 6 concludes.

2. Quality of Forecast Learning in a Stylized Factor-Augmented Model

Retail companies use proprietary forecasting models, which are often the result of a complicated engineering effort, tailored to the particularities of the business. Theoretical properties of such complicated forecasting models may never be understood completely. What we strive to do here is to give a theoretical benchmark based on a well-known, state-of-the-art model used in academia to model panel time series and forecasting. In particular, we utilize the Augmented Factor model; see, for example, Bai and Ng (2002), Bernanke et al (2004), and Bai (2009). The Augmented Factor model is a rather general model, encompassing the standard fixed effects model from microeconometrics, factor models from macro, finance, and matrix completion models, and traditional regression models. Within this fairly general, yet tractable model, we ask the question of how forecast errors would be determined and what role do the data size, the number of products and number of time periods in the available history, have. We also pose the same question at a more general level, when we consider various deviations and generalization of the model, still reaching similar conclusions. We think of this section as providing independent, theoretical evidence on the likely causal effects of big data on the quality of demand forecasting. We also use this section to inform our empirical analysis.

2.1. **Theoretical Framework and Questions.**

Our goal is to perform comparative statics on the quality of demand learning by the retailer firm with respect to the size of the data that it obtains over a period of time of operating in the retail business.

An important challenge in modelling demand for retailers is that demand may have a non-trivial non-stationary component; especially for rapidly growing retailers. Hence, we first consider a simple, multiplicative model of the following type:

M1: The quantity $Q_{i,t}^k$ sold by the retailer k of a product i at time t, obeys the following equation:

$$\underbrace{Q_{i,t}^k + 1}_{\text{quantity}} = \underbrace{V_{i,t}^k}_{\text{velocity multiplier}} \cdot \underbrace{Q_{i,t}^0}_{\text{base demand index}},$$

$$t = 1, ..., T; \quad i = 1, ..., N;$$

where $V_{i,t} \geq 1$ is a known velocity variable, which describes the stochastic, possibly non-stationary level of the series, and $Q_{i,t}^0 \geq 1$ is a reference demand level, where $\{Q_{i,t}^0\}_{t=0}^{\infty}$ is stationary for each i. Note that

$$Q_{i,t}^k = 0 \text{ if and only if } V_{i,t}^k = 1 \text{ and } Q_{i,t}^0 = 1.$$

In model M1, the quantity $Q_{i,t}$ (plus 1) is determined by the base demand times the multiplier $V_{i,t}$ that captures the "size of the firm" in product i sales as well as "velocity" of the product:

- Velocity $V_{i,t}$ reflects the notional popularity of the product i at time t, and represents the product-specific size of the retailer, as specific to the product. A simple example of $V_{i,t}$ is given by the lagged sales times a growth rate

$$V_{i,t}^k = (Q_{i,t-1}^k + 1).$$

  We can normalize the velocity at $V_{it}^k = 1$ for $t = 1$ for retailer $k = 1$. The vector of velocities $\{V_{i,t}\}_{i=1}^N$ characterizes the overall "size of the retailer".
- Both velocity and the base demand are conceptual quantities, which we introduce to perform the comparative statics on the quality of demand learning by the retailer firm with respect to the size of the data. We will discuss the ramifications of the assumption that the velocity part of demand is "known" below.

We next give a model for the base demand level, following the Augmented Factor panel data model introduced above. In this model, both time series dimension T and cross-sectional dimensions N play a crucial role in learning.

M2. The base demand of a product i at time t, obeys the following equation:

$$\underbrace{Q_{i,t}^0}_{\text{base demand index}} := \exp\left(\underbrace{\alpha_i' F_t}_{\text{latent factors}} + \underbrace{X_{i,t}'\beta}_{\text{observed factors}} + \underbrace{\epsilon_{i,t}}_{\text{stochastic shocks}}\right),$$

$$t = 1, ..., T; \quad i = 1, ..., N;$$

where the latent shocks $\{\epsilon_{i,t}\}_{t=1}^{\infty}$ obey

$$\epsilon_{i,t} \mid \alpha_i, \{\epsilon_{i,l}\}_{l=l}^{t-1}, \{F_l\}_{l=1}^{t}, \{X_{i,l}\}_{l=1}^{t}, \{V_{i,l}^k\}_{l=1}^{t} \sim N(0, \sigma_i^2).$$

Vector F_t contains 1 as the first component and obeys the normalization $\mathrm{E}F_t F_t' = I$ and follow the transition equation

$$F_t = \Phi_1 F_{t-1} + ... + \Phi_d F_{t-d} + \nu_t, \tag{2.1}$$

where $v_t \sim N(0, \Sigma^v)$ are i.i.d. shocks across time, and such that $\{F_t\}$ is stationary. The latent loadings $(\alpha_i)_{i=1}^N$ and shock sizes $(\sigma^i)_{i=1}^N$are identically distributed, obeying normalization $\mathrm{E}\alpha_i\alpha_i' = \mathrm{Diagonal}$ and $\max_i \sigma_i \leq L$. The relevant moments are bounded, namely F_t and $X_{i,t}$ are sub-exponential with common upper bound on scale L, uniformly for all i and t.

In model M2, the base quantity index $Q_{i,t}^0$ is determined by latent factor components plus observed components plus stochastic shocks:

- Time-varying factors F_t are latent time-varying common factors, such as macro-economic factors, seasonality, and fashion factors.
- The product varying factors α_i are the product-specific loadings on the latent factors F_t, to which product demand responds differently. The include the conventional fixed effects model, when $F_t = 1$.
- The observed component is determined by a p-dimensional vector $X_{i,t}$ of observed time-varying product features, such as prices of the product as well as its substitutes and complements, multiplied by a common parameter β.
- The latent shocks $\epsilon_{i,t}$ are unobserved, unlearnable error components.

2.2. **Comparative Statics for Quality of Learning.** Within the model posed above, we ask the following comparative statics questions:

Q1: How does N, the number of products, and T, the number of available time periods, affect the quality of learning/demand forecasting? Are there decreasing returns to scale with respect to both N and T?

Q2: How does N affect the quality of learning, where adding new products changes velocity of other products? For example, how does adding many substitutes for an item i, which lowers velocity $V_{i,t}$, affect the quality of learning with respect to the data size (N, T)?

Q3: How does velocity/size of the company affect the quality of learning? For example, for a big and a small retailer, with velocities of the first greater than the velocities of the second, $\{V_{i,t}^1\}_{i=1}^N > \{V_{i,t}^2\}_{i=1}^N$, $t = 1, ..., T$, are there different qualities of learning with respect to the data size (N, T)?

Let $\hat{Q}_{i,t}^k$ denote the forecast of the demand $Q_{i,t}^k$ made by forecaster k, taking the form:

$$\hat{Q}_{i,t}^k + 1 = V_{i,t}\hat{Q}_{i,t}^{0,k},$$

where $\hat{Q}_{i,t}^{0,k}$ denotes the forecast of the base demand index $Q_{k,t}^0$ made by forecaster k.

We next study the quality of forecast in relative terms:

$$\text{relative error}_{i,t}^k := \frac{|(\hat{Q}_{i,t}^k + 1) - (Q_{i,t}^k + 1)|}{(Q_{i,t}^k + 1)},$$

and also in absolute terms

$$\text{absolute error}_{i,t}^k := |\hat{Q}_{i,t}^k - Q_{i,t}^k|.$$

Within the model M.1, we immediately arrive at the following obvious assertion.

ASSERTION 1: *The relative forecast error does not depend on velocity $V_{i,t}^k$, it only depends on the relative error of forecasting base demand $Q_{i,t}^0$ with $\hat{Q}_{i,t}^{0,k}$:*

$$\textit{relative error}_{i,t}^k = \frac{|\hat{Q}_{i,t}^{0,k} - Q_{i,t}^0|}{Q_{i,t}^0}.$$

The absolute error is an increasing function in velocity $V_{i,t}^k$:

$$\textit{absolute error}_{i,t}^k = V_{i,t}^k |\hat{Q}_{i,t}^{0,k} - Q_{i,t}^0|,$$

holding the base demand and the base forecast fixed, at $\hat{Q}_{i,t}^{0,k} \neq Q_{i,t}^0$.

This admittedly trivial assertion on velocity runs counter to the following intuitive statement: "the bigger the sales of a product, the more data we have, the more accurate the forecast we should have." Systematically higher sales are captured by the velocity multiplier $V_{i,t}$, and V_{it} simply cancels out in the definition of the relative error, as we have seen above. What about absolute error? Higher velocity leads to a higher absolute forecasting error, and lower velocity leads to lower forecasting error. Indeed, in the extreme case, it is very easy to forecast demand for products that do not sell. Does velocity affect the quality of learning the base demand? We examine this question below.

In the model above, one can estimate the base demand as follows. Since velocity is known, dividing through by velocity and taking logs, yields the regression model:

$$\log((Q_{i,t}^k + 1)/V_{i,t}^k) = \alpha_i' F_t + X_{i,t}'\beta + \epsilon_{i,t}, \quad i = 1, ..., N, \quad t = 1, ..., T. \tag{2.2}$$

We see immediately that in this model velocity will play *no role* in learning the parameters of the base forecast, and we only need to analyze the impact of data size (N, T) on the quality of the base forecast. The base demand can be estimated usually at the rate $1/\sqrt{T} + 1/\sqrt{N}$ in this model, and this is what we will show below.

Basics of Base Demand Forecasting. In what follows we go though the basics of the base demand forecasting. We emphasize here the conceptual blocks that are simple and useful for empirical economists.

Consider the least squares estimator

$$(\{\hat{\alpha}_i\}_{i=1}^N, \{\hat{F}_t\}_{t=1}^T, \hat{\beta}),$$

which solves

$$\min_{(\{\alpha_i\}_{i=1}^N, \{F_t\}_{t=1}^T, \beta) \in \Theta_{N,T}} \sum_{i=1}^N \sum_{t=1}^T (\log((Q_{i,t}^k + 1)/V_{i,t}^k) - \alpha_i' F_t - X_{i,t}' \beta)^2,$$

where

$$\Theta_{N,T} = \left\{ \{\alpha_i\}_{i=1}^N, \{F_t\}_{t=1}^T, \beta : \quad \frac{1}{N} \sum_{i=1}^N \alpha_i \alpha_i' = \text{diagonal}; \quad \frac{1}{T} \sum_{t=1}^T F_t F_t' = I \right\}.$$

Restrictions on the parameter space reflect the normalization assumptions on the latent factors, requiring them to be orthogonal to each other; see, e.g., Bai (2009) for the detailed discussion. The time-varying factors are normalized to have unit variance, while the product loadings have unrestricted variance.

It follows from Bai (2009)'s analysis, that under the condition M2 and additional regularity conditions, as $N, T \to \infty$, the estimator obeys for each t and i (that is, pointwise),

$$\begin{array}{ll} \|\hat{F}_t - F_t\| = O_p(1/\sqrt{N}), & \|\hat{\alpha}_i - \alpha_i\| = O_p(1/\sqrt{T}), \\ |\hat{\alpha}_i' \hat{F}_t - \alpha_i' F_t| = O_p(1/\sqrt{T} + 1/\sqrt{N}), & \|\hat{\beta} - \beta\| = O_p(1/\sqrt{T} + 1/\sqrt{N}), \end{array} \tag{2.3}$$

with the average squared errors bounded as:

$$\frac{1}{T} \sum_{t=1}^T \|\hat{F}_t - F_t\|^2 = O_p(1/N), \quad \frac{1}{N} \sum_{t=1}^N \|\hat{\alpha}_i - \alpha_i\|^2 = O_p(1/T). \tag{2.4}$$

Instead of stating the additional conditions here, we will simply assume the performance bound (2.3-2.4).

The performance bound has an intuitive interpretation:

- even if we observe directly the product specific loading parameters α_i, we need sufficiently many cross-sectional observations N to learn the time-varying latent factors F_t, with error scaling like $1/\sqrt{N}$, and
- even if we observe the latent factors F_t, we need sufficiently many time series observations to learn the product specific parameters α_i, with error scaling like $1/\sqrt{T}$.

Taken altogether the rate of learning the inner product $\alpha_i' F_t$ is

$$1/\sqrt{T} + 1/\sqrt{N}.$$

Note that the worst case rate for learning common parameter β is also $1/\sqrt{T} + 1/\sqrt{N}$. However, the common parameters can be learned at the faster rate of $1/\sqrt{TN}$ under the additional side conditions (for example, $T \propto N$). Even if such a faster rate is available for the common component, it will not determine the rate of convergence of the overall forecast.

Moreover, for prediction purposes, we need to estimate the parameters of the autoregressive model for latent factors. We can obtain the estimator $(\hat{\Phi}_l)_{l=1}^d$ as the solution of the least squares problem:

$$\min_{(\Phi_l)_{l=1}^d} \sum_{t=d+1}^{T} (\hat{F}_t - \Phi_1 \hat{F}_{t-1} - ... - \Phi_d \hat{F}_{t-d})^2.$$

Under condition M2 and under (2.3-2.4), it can be shown that the estimator obeys

$$\|(\hat{\Phi}_l)_{l=1}^d - (\Phi_l)_{l=1}^d\| = O_p(1/\sqrt{N} + 1/\sqrt{T}). \tag{2.5}$$

It is reasonable to assume that retailers' forecasts can not be worse than using this estimator. Since the above rates can not be improved in general, it is also reasonable to assume that the rates achieved by the retailers' forecasters can not be better than the rates above. Given, this, without loss of generality for the rate of learning results that follow, we assume that the retailers use the least squares estimators above:

L. The forecasting team of retailer k use the correct model, which they know up to the parameters, and they know velocity $V_{i,t}^k$ at time t. The latent factors are treated as unknown parameters. They follow a good practice, namely they use estimators that exhibit performance bounds (2.3)-(2.5). Without loss of generality for the rate of learning results, assume that they use the estimators used above.

We next turn to forecasting. The base *oracle forecast* is given by:

$$\hat{Q}_{i,t}^{0,oracle} = \exp\Big\{\alpha_i'\bar{F}_t + X_{i,t}'\beta + a_i\Big\},$$

$$\bar{F}_t = \Phi_1 F_{t-1} + ... + \Phi_k F_{t-k},$$

where a_i is a constant chosen depending on the loss function used to evaluate the forecast:

$$a_i = (\sigma_i^2 + \alpha_i'\Sigma_v\alpha_i)/2, \quad \text{if predicting mean},$$

$$a_i = \sqrt{(\sigma_i^2 + \alpha_i'\Sigma_v\alpha_i)}\Phi^{-1}(p), \quad \text{if predicting } p\text{-th quantile},$$

That is, we define the *oracle forecast* for this model as the forecast one would make in the presence of unlimited data to estimate the parameters $(\{\alpha_i\}_{i=1}^N, \{F_t\}_{t=1}^T, \beta, \{a_i\}_{i=1}^N)$ without error.

In reality, however, the parameters are not known to the forecasters, and have to be estimated. Using the estimator described above, one can form a feasible one-step ahead base forecast for $t = T + 1$ using data of size (T, N), given by:

$$\hat{Q}_{i,t}^{0} = \exp\Big\{\hat{\alpha}_i'\widehat{\bar{F}}_t + X_{i,t}'\hat{\beta} + \hat{a}_i\Big\},$$

where $\widehat{\bar{F}}_t = \hat{\Phi}_1\hat{F}_{t-1} + ... + \hat{\Phi}_k\hat{F}_{t-k}$. Moreover, the estimator of a_i is defined by the plug-in principle (see Appendix A).

Dependency of Quality of Base Forecast Learning on Data Size. The basic calculations above imply the following assertion, derived in the appendix.

ASSERTION 2. (Dependency of Quality of Base Forecast Learning on Data Size). *Suppose conditions M1-2 and L hold, then the relative forecast error does not depend on the velocity. Define also:*

$$\textit{irreducible error}_{i,t} := \frac{|\hat{Q}_{i,t}^{0,oracle} - Q_{i,t}^{0}|}{Q_{i,t}^{0}},$$

which is the forecasting error one would make in in the presence of unlimited data to estimate the parameters $(\{\alpha_i\}_{i=1}^N, \{F_t\}_{t=1}^T, \beta, \{a_i\}_{i=1}^N)$ *without error. Letting* K *be a positive constant, there are* N *and* T *sufficiently large, such that the relative forecast error is bounded in expectation as follows: for* $t = T + 1$*,*

$$\begin{aligned} \mathrm{E}|\text{relative error}_{i,t}^{k}| \wedge K &\leq \mathrm{E}|\text{irreducible error}_{i,t}| \\ &+ C_i(\sqrt{1/N} + \sqrt{1/T}), \end{aligned}$$

where $C_i \leq C$*, and* C *is a constant that does not depend on* N *and* T*. Moreover, the probability that the relative error exceeds a threshold* c *is given by:*

$$\Pr(|\text{relative error}_{i,t}^{k}| > c) \leq \Lambda_i + C_i(\sqrt{1/N} + \sqrt{1/T}),$$

where

$$\Lambda_i = \mathrm{P}(|\text{irreducible error}_{i,t}| > c/2),$$

where $C_i \leq C$*, and* C *is a constant that depends on* c *but does not depend on* N *and* T*.*

For the detailed argument, please see Appendix A. The relative forecast error is bounded by a term proportional to $\sqrt{1/N} + \sqrt{1/T}$ and a constant term equal to the expectation of the irreducible error, which is the amount of error that would result from the (infeasible) oracle forecast formed with unlimited data.

We state the main consequences on comparative statistics.

IMPLICATION 1 (COMPARATIVE STATICS ON THE QUALITY OF FORECAST LEARNING). *Suppose conditions M1-2 and L hold. We have the following answers to the questions Q1-Q3.*

A1: *There are diminishing improvements in quality of forecast with respect to both N and T, with the relative error decreasing to the irreducible size at the rate $1/\sqrt{N} + 1\sqrt{T}$.*

A2: *Velocity $V_{i,t}$ has no impact on the relative forecast error, so if the fraction of products with higher velocity goes up, without affecting N and T and average (across products) error of forecasting base demand indices, the average (across products) quality of overall forecast stays constant in relative terms.*

A3: *Retailers with different vectors of velocities (retail sizes) and the same data size (N, T) make the relative forecast errors that decay to the same irreducible size at the same rate $1/\sqrt{N} + 1/\sqrt{T}$.*

We next make an observation about the *absolute error*:

$$\text{absolute error}^k_{i,t} := V_{it}|\exp(\hat{\alpha}_i'\hat{F}_t + X_{i,t}'\hat{\beta} + \hat{a}_i) - \exp(\alpha_i' F_t + X_{i,t}'\beta + \epsilon_{i,t} + a_i)|.$$

Implication 2 (Comparative Statics on The Quality of Forecast Learning). *Suppose conditions M1-2 and L hold, we have the following supplemental answers to the questions Q2-Q3.*

A2′: *Velocity $V_{i,t}$ increases the absolute forecast error, so if the fraction of products with higher velocity goes up, without affecting N and T and average (across products) relative error of the forecast, the average (across products) quality of forecast becomes worse in absolute terms.*

A3′: *Retailers with higher vector of velocities (retail sizes), having the same forecast for base demand, incur higher absolute forecast error.*

2.2.1. *Inactive Products.* A significant fraction of products are never sold. We note that this is a special case where the base demand can be learned very fast, faster than $1/\sqrt{T} + 1/\sqrt{N}$. More formally, we can call a product *inactive* if the demand is zero most of the time, or,

$$\log((Q^k_{i,t} + 1)/V^k_{i,t}) = 0, \text{ for all } t = 1, \ldots, T.$$

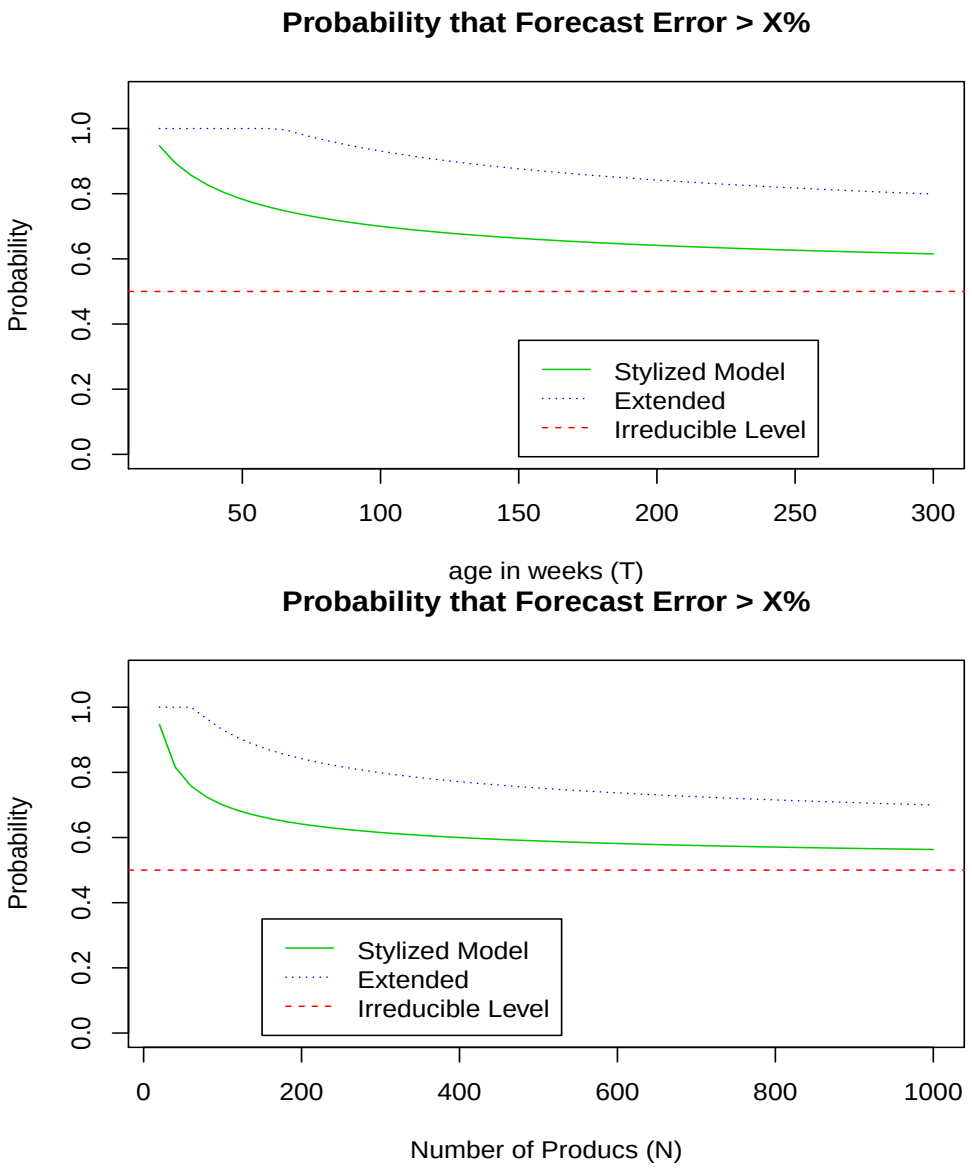


FIGURE 1. A schematic impact of N and T on the Quality of Learning in the Stylized Model M.1-2 and in the Extended Model described in the next subsection, in equation (2.7) using parameters $a = b = 1/3$. In both models, there are diminishing improvements in the quality of forecast as N and T increase, and the forecast error can not decrease below the irreducible level. In the Extended Model, which is a more realistic approximation, the improvements occur even more slowly, with data size helping much less.

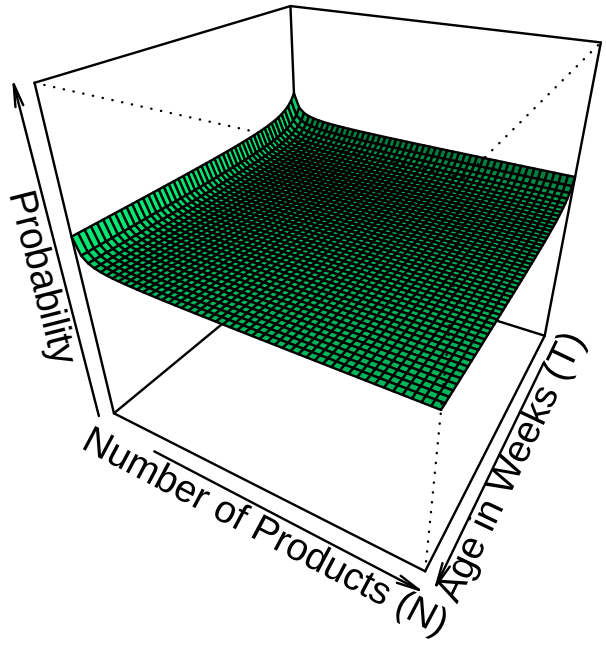


FIGURE 2. The Joint Impact of N and T on the Quality of Learning in the Stylized Model M.1-2. There are diminishing improvements in the quality of forecast as N and T increase.

In the model that we wrote down, these zeroes can be learned very fast by simply testing if $\log((Q_{i,t}^k + 1)/V_{i,t}^k) = 0$ for $t = 1, ..., T$ and the probability of making a mistake goes to zero exponentially fast. This means that very little data is needed to forecast in-active products.

ASSERTION 3. (Informal: Dependency of Quality of Forecast Learning on Data Size for In-Active Products). *For in-active products, we can state the following informal approximation, which may be sharper:*

$$\Lambda_i \approx 0, \quad C_i \approx 0,$$

once N and T are sufficiently large.

In Appendix B we state some empirical results that seem to offer some empirical evidence supporting this theoretical prediction.

2.3. **Comparative Statics under Extensions and Complications.** We also discuss extensions and complications that arise beyond the stylized model:

E1: Velocities $V_{i,t}$ are the main source of non-stationarity in the problem. Suppose that velocities are not known, but unmodeled changes in velocities occur slowly enough, to allow for local learning of the best forecast, which implicitly introduces a model with time-varying parameters. How does this affect the quality of learning with respect to data size (N, T)?

E2: The production model could involve a much greater degree of complexity than the augmented factor linear model specified above. For example, the complexity/dimension of the model p and k could be increasing with the size of the data. Or the model could be entirely non-parametric. How does this affect the quality of learning with respect to data size (N, T)?

Let us try to analyze E.1 from an informal point of view. Suppose we use a proxy for velocity given by $P_{i,t}$ and that this proxy is not perfect, then defining

$$a_{i,t}^k = \log(P_{i,t}^k / V_{i,t}^k)$$

the model for predicting base forecast becomes:

$$\log((Q_{i,t}^k + 1)/P_{i,t}^k) = a_{i,t}^k + \alpha_i' F_t + X_{i,t}'\beta + \epsilon_{i,t}, \quad i = 1, ..., N, \quad t = 1, ..., T. \tag{2.6}$$

The term $a_{i,t}^k$ is potentially non-stationary and correlated with latent factors and observed features. Since $a_{i,t}^k$ can not be treated as an orthogonal error, any model that omits $a_{i,t}^k$ is subject to the (time changing) specification error. That is, the best approximating model

$$\underbrace{\tilde{\alpha}_i' \tilde{F}_t + X_{i,t}'\tilde{\beta}}_{\text{best approximating model}} \approx a_{i,t}^k + \alpha_i' F_t + X_{i,t}'\beta$$

will depend on the historical window over which t varies. In order to construct the best approximating model that is most pertinent for the most recent history, it is therefore common to limit the observation data only to some recent history:

$$T - M \leq t \leq T.$$

This limits the effective amount of data and makes the effective time series dimension M and not T. Under this scenario we expect that the impact of large T on the quality of forecast is in fact limited by

$$1/\sqrt{\min(M, T)}.$$

The marginal effect of data size T here is zero, once $T > M$, and so the marginal effect of data size is always lower in this extended model than in the stylized model. Surprisingly this also limits the usefulness of N, since for any fixed constant C,

$$1/\sqrt{CN} + 1/\sqrt{\min(M, T)} \geq 1/\sqrt{\min(CN, \min(M, T))} = 1/\sqrt{\min(M, T, CN)}$$

In summary, *non-stationarity severely limits the usefulness* of both N and T.

For the second scenario we expect the results to change as follows. If d is the number of latent factors, and p is the number of covariates, which are modeled as large, non-negligible compared to N or T, we expect the following change in the performance bound

$$\Pr(|\text{relative error}_{i,t}^k| > c) \leq \Lambda_i + C_i(\sqrt{d/N} + \sqrt{d/T} + \sqrt{p/TN}).$$

Hence the errors now become much larger in magnitude, while decreasing to zero at the slower rates that depend on how d and p depend on N and T. The marginal effects of data sizes can be higher in these models than in the stylized models (depending on the constants).

There are other extensions we can consider, all leading to deterioration of quality of learning, making errors larger in magnitude, resulting in slower than $\sqrt{N}$ and $\sqrt{T}$ learning rates:

$$\Pr(|\text{relative error}_{i,t}^k| > c) \leq \Lambda_i + C_i(N^{-a} + T^{-b}), \tag{2.7}$$

with $0 \leq a, b \leq 1/2$. For example, in the model where all parameters of the base demand are heterogeneous across products, cross-sectional dimension can play no role, leading to $a = 0$ and the learning exploits the time series variation only. All kinds of nonparametric learning of base demand also leads to slower than $1/\sqrt{T}$ rates and $1/\sqrt{N}$ rate. We visualize the impact on the quality of learning in Figure 1. The marginal effect of N in this model is $-C_i a N^{-a-1}$, which can be larger than the marginal effect $-C_i(1/2)N^{-3/2}$ in the stylized model (this comparison depends on the constants C_i that can differ in the two places).

In summary, the marginal value of each data point can be either higher or lower in these extended models, and we can only try to determine this value empirically. The bottom line, however, regardless of the extension discussed here, there are diminishing returns to data sizes in each model, with marginal value of each data point decreasing to zero as the data sizes increase.

3. Empirical Specification and Results for the Case of Forecasting Demand for Electronics

3.1. **Data and Empirical Specification.** We will now perform tests of the comparative statics derived in the previous section using a panel data set that consists of up to $H = 234$ weekly observations on $M = 6079$ products, which is a random sample of traded products from the electronics product group.[1] The panel covers weekly observations for the period ranging from 2012/12/08 to 2017/06/03.

We observe $Q_{i,t}$, the quantity of product i sold during week

$$t = 1, \ldots, H,$$

as well the corresponding one-week ahead mean forecast $\hat{Q}_{i,t}$ that was produced using data available at one week ago, that is, at $t-1$. We also shall use the following variables:

[1]Note that the relevant dimensions of panel data here are H and M – these symbols are used in order not to confuse with the symbols "N" and "T" used in the previous section and here, which have a related, but different meaning.

- $T_{i,t}$ = the "Age" of product at the company, which describes the length of data available for product i at time t, as consumed by the forecasting model. The own history data, especially if available for substantive ranges, can be used to build a high quality forecast.
- $N_{i,t}$ = the number of products in the same category ("NCat") as product i at time $t-1$, that was consumed by the forecasting model and for which the forecasts were produced.[2] It is reasonable to expect, from both theory and common sense, that data on these products can be used to capture seasonality and fashion patterns, improving the forecast based upon own history alone.

We define the relative forecast error as before, namely as

$$|Q_{i,t} - \hat{Q}_{i,t}|/(Q_{i,t}+1).$$

We would like to assess how the size of the available data $(T_{i,t}, N_{i,t})$ affect the relative forecast error.

Specifically, motivated by the previous comparative statics results we define the following dependent variable

$$Y_{i,t} = 1\{\underbrace{|\hat{Q}_{i,t} - Q_{i,t}|/(Q_{i,t}+1) > X}_{\text{"Big Forecast Error Event"}}\}.$$

This variable describes the event of making a *big forecast error*, namely $X\%$ error. The threshold of X is chosen so that this threshold is exceeded for any random product-week pair (t, i) with probability $P = 30\%$. We obtained quantitatively analogous results with larger P's. When we extend the analysis to all other product groups in Section 4, the same uniform rule for $P = 30\%$ defining the big error event will be applied.

[2]The categories include Home Entertainment, Home Audio, Portable Electronics, Portable Digital Players, GPS and Car Audio, PC Products, Audio Speakers, Audio Receivers, Wireless Audio, Audio Components, Warranty & Services, Portable Media Players, Entertainment Software - Sports & Outdoors, Headphones,Accessories Power, Video Components, Cables, Other Accessories Software - Business & Productivity, Uncategorized, as well as several other deprecated categories.

Some basic descriptive statistics for the variables defined above is given in the following table.

TABLE 1. Descriptive Statistics

Statistic	N	Pctl(25)	Pctl(75)	Mean	Max	Min
BigErrorEvent	497,078	0	1	0.300	1	0
Age	497,078	68	187	127.950	278	0
NCat	497,078	4,087	10,437	8,083.528	15,705	1

Figure 3 shows the cross-sectional averages of Relative Forecast Error, Big Error Event occurrence, Age, and Number of Products, for each week $t = 1, ..., H$. Figure Figure 4 shows the fraction of products with no sale by week and fraction of the forecasted products with no sale. Note that that a big improvement in the Relative Forecast Error occurring at about two-third of the time window is due to drastic addition of the new products for which the forecast is made at about the same time.

Giving the raw data, it is useful to ask an "identification question": what is the source of variation that will help us identify the projection coefficients in the predictive models? Figure 5 shows the source of variation: after taking out the time and product fixed effects from the key variables that measure the data sizes, we are left with residuals that exhibit variation. The Number of Products have considerable independent variation left after taking out the product and time fixed effects: Number of Products $N_{i,t}$ varies across categories, and the time effects can only take out the aggregate variation. The Age of Products $T_{i,t}$ varies across (i, t) as well, but note that the amount of variation is small. The small variation still allows us to pin down the projection coefficients on age $T_{i,t}$ in the predictive models that we will estimate. We note that identification of projection coefficients does not necessarily imply identification of causal impacts, as predictive models that we will estimate have a causal interpretation only under rather strong exogeneity conditions.

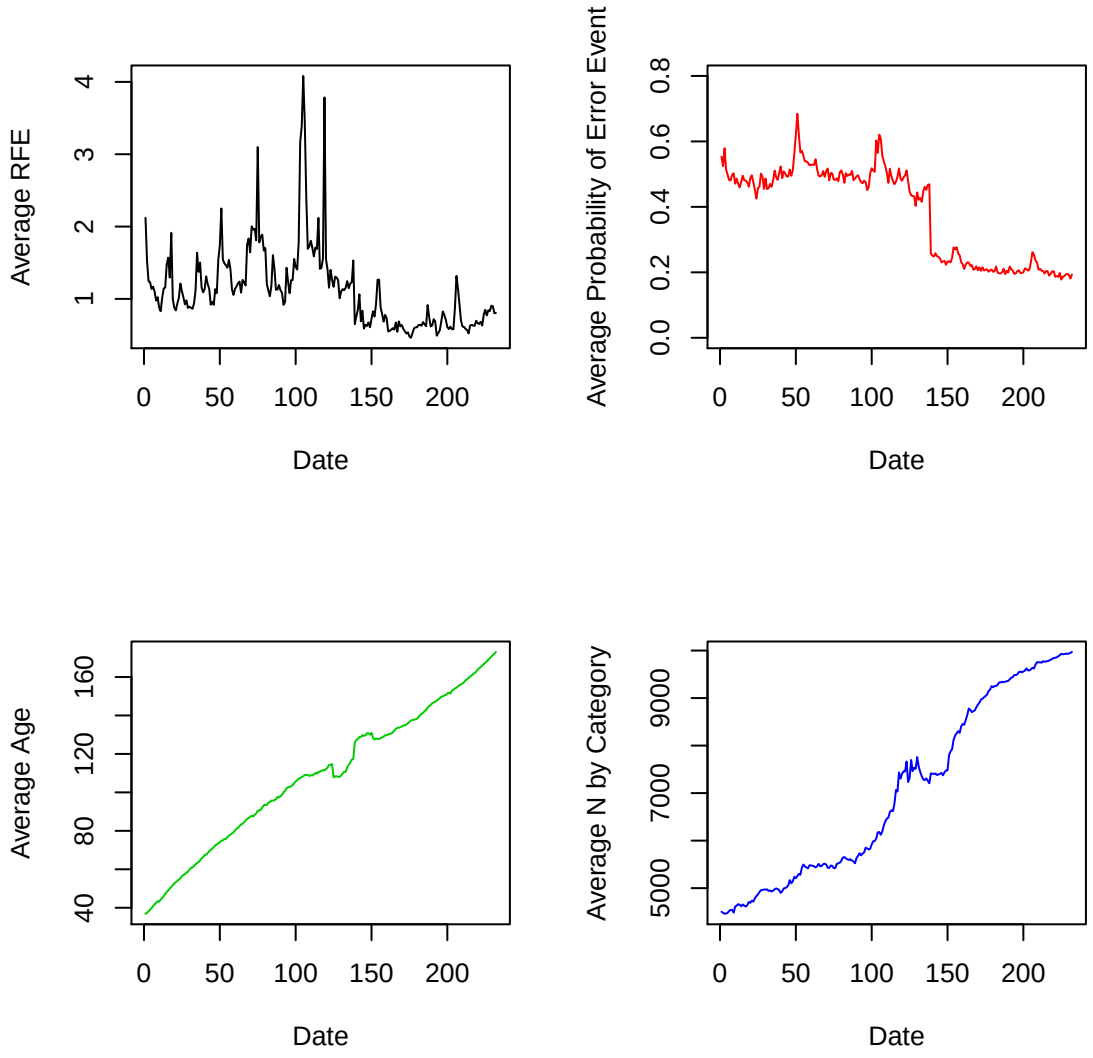


FIGURE 3. Cross-sectional averages of Relative Forecast Error, Big Error Event occurrence , Age, and Number of Products, by Date

Theory-Motivated Empirical Specification. We first consider the following linear prediction equations for $Y_{i,t}$:

$$\begin{aligned} Y_{i,t} &= \underbrace{\alpha_i + \beta_t}_{\text{Product and Time Effects}} \\ &+ \underbrace{\delta_1 1(T_{i,t} > 20) + \delta_2 (T_{i,t} > 20)/\sqrt{T_{i,t}} + \delta_3 1(T_{i,t} > 20)/T_{i,t}}_{\text{Product Age Effect}} \\ &+ \underbrace{\gamma_1 1(N_{i,t} > 200) + \gamma_2 1(N_{i,t} > 200)/\sqrt{N_{i,t}} + \gamma_3 1(N_{i,t} > 200)/N_{i,t}}_{\text{Number of Products Effect}} \\ &+ \epsilon_{i,t}, \end{aligned}$$

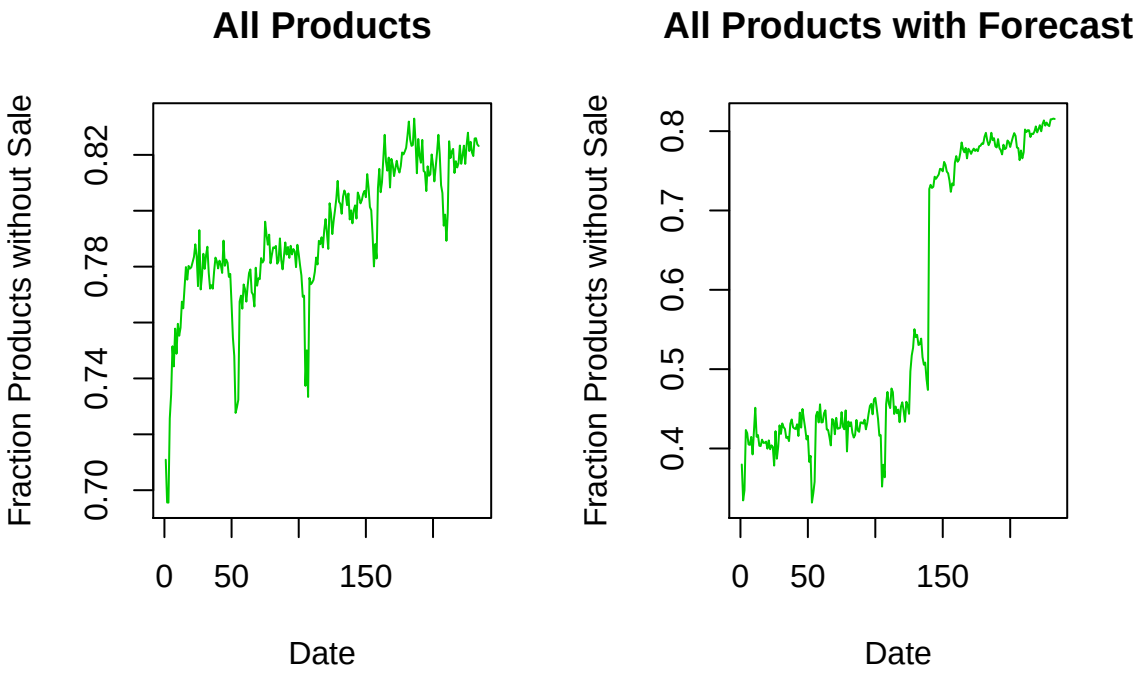


FIGURE 4. Fraction of products with no sale by week , and fraction of forecasted products with no sale.

where the terms α_i are the product specific fixed effects, which capture the notion that some products are inherently harder to forecast than others, even using the same data size $(N_{i,t}, T_{i,t}) = (N, T)$. The terms β_t are the time effects, which capture that on some dates it may be harder or easier to forecast than on others; moreover, these time effects also partly capture the overall improvement in the forecasting model that produces $\hat{Q}_{i,t}$. We consider three specifications for time effects:

(1) No Trend: the specification with $\beta_t = 0$ for all $t = 1, ..., H$.
(2) Smooth Trend: the specification, where for some parameters b_1 and b_2, $\beta_t = b_1(t/H) + b_2(t/H)^2$.
(3) Time Fixed Effects: the specification, where β_t's are unrestricted parameters.

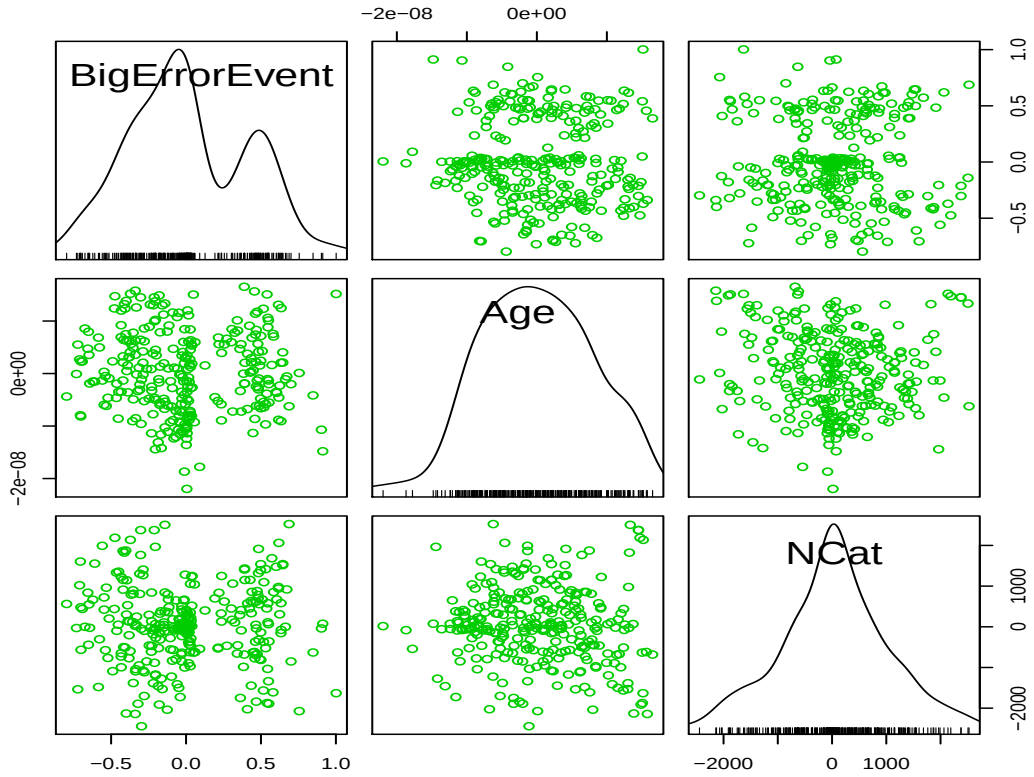


FIGURE 5. A Scatter Plot of 300 randomly selected data points on Age, Number of Products, and Big Error Event Indicators, after taking out product and time effects. The Number of Products has considerable independent variation left after taking out the product and time fixed effects, since $N_{i,t}$ varies across categories, and the time effects can only take out the aggregate variation. The Age of Products $T_{i,t}$ varies across (i,t) as well, but the amount of variation left after taking our the product and time effects is small. The small variation still allows us to pin down the projection coefficients on age $T_{i,t}$ in the predictive models that we will estimate.

The term "Product Age Effects", or the "T Effect", is meant to capture the nonlinear effect of the age of product i on the relative forecast error. The leading part of this term, $\delta_1 1(T_{i,t} > 20) + \delta_2 (T_{i,t} > 20)/\sqrt{T_{i,t}}$, is motivated by the theoretical

bounds on the relative forecasting error derived in the theoretical stylized model. As above, the remaining term $\delta_3 1(T_{i,t} > 20)/T_{i,t}$ is meant to capture deviations and nonlinearities relative to the leading term. For example, this term could capture the "moderation" of the impact of data size due to unknown velocity multiplier (if the coefficient $\delta_3 > 0$), as discussed in Section 2.3 on the extended comparative statics.

From the discussion in Section 2 we expect that the coefficients will exhibit the following signs:

$$\delta_1 \leq 0 \text{ and } \delta_2 \geq 0 \text{ and } \delta_3 \leq 0,$$

although we will not restrict the signs in the estimation. The expected sign $\delta_1 \leq 0$ is obvious, while to make sense of $\delta_2 \geq 0$, note that we expect that the derivative of $\delta_2(T_{i,t} > 20)/\sqrt{T_{i,t}}$ w.r.t. $T_{i,t}$ is non-positive, which translates into $\delta_2 \geq 0$. Similarly, if the higher order term is meant to moderate the effect the of the first-order term, we expect the derivative of $\delta_3 1(T_{i,t} > 20)/T_{i,t}$ with respect to $T_{i,t}$ to be non-negative, which translates into $\delta_3 \leq 0$.

Similarly, the term "Number of Products Effect", or the "N Effect", is meant to capture the nonlinear effect of the number of products $N_{i,t}$ in a similar broad category as i, on the relative forecast error. The leading part of this term, $\gamma_1 1(N_{i,t} > 200) + \gamma_2 1(N_{i,t} > 200)/\sqrt{N_{i,t}}$, is motivated by the theoretical bounds on the relative forecasting error derived in the stylized model applied to i and products in the same category. And the remaining part $\gamma_3 1(N_{i,t} > 200)/N_{i,t}$ is the square of $1(N_{i,t} > 200)/\sqrt{N_{i,t}}$ and is meant to capture deviations and nonlinearities, such as "moderation" or "acceleration" effects depending on the sign of the coefficient, relative to the leading term.

From the discussion in Section 2 we could expect that the coefficients will exhibit the following signs:

$$\gamma_1 \leq 0 \text{ and } \gamma_2 \geq 0 \text{ and } \gamma_3 \leq 0,$$

although we will not restrict the signs in the estimation, as stated above.

The projection error $\epsilon_{i,t}$ is by definition orthogonal to the space spanned by all variables, including the product and time fixed effects.

We believe it is reasonable to always include product fixed effects as the leading empirical specification. Indeed, the product fixed effects control for changing product mix and for the fact that the products are added in a non-random manner. Time effects capture phenomena like forecasting demand in holiday period might be more difficult than in non-holiday periods as well as roll-outs of upgraded forecasting model over time. The threshold of $T_{i,t} > 20$ in the definition above is meant to signify a minimum sample size for time series dimension such that it is conceivable that the theoretical motivations of the previous section apply here. Likewise, $N_{i,t} > 200$ was chosen similarly, based upon our practical experience that substantial cross-sectional size is needed to start to learn factors (e.g. seasonality) properly. Third, our model as well as the model described below are predictive models, they describe how changes in the data sizes $(N_{i,t}, T_{i,t})$ affect the predicted probability of the large forecast errors. We don't necessarily ascribe a causal interpretation to these results. [3]

Agnostic Empirical Specification. We also consider a flexible, easy-to-interpret model that does not impose the functional form of the theoretical model, and allows the data to speak for itself more strongly. The model is still informed by theory in that we are linking the probability of big error event to the data sizes $T_{i,t}$ and $N_{i,t}$, but we relax the functional form so that the empirical results can more easily differ from the predictions derived from the theoretical model.

[3]The models can have a causal interpretation under some well-known exogeneity conditions. Getting a causal interpretation out of the fixed effects model is implausible in the setting with short time-series dimension.

In this agnostic specification, we divide the ranges of $T_{i,t}$ and $N_{i,t}$ into regions and predict $Y_{i,t}$ using region dummies as well as product and time effects:

$$Y_{i,t} = \underbrace{\alpha_i + \beta_t}_{\text{Product and Time Effects}} + \underbrace{\sum_{j=1}^{5} \delta_j 1(T_{i,t} \in A_j)}_{\text{Product Age Effect}} + \underbrace{\sum_{k=1}^{5} \gamma_k 1(N_{i,t} \in B_k)}_{\text{Number of Products Effect}} + \epsilon_{i,t},$$

where the projection error $\epsilon_{i,t}$ is by definition taken to be orthogonal to the space spanned by all variables, including the product fixed effects and time effects. Here we consider three specifications for the time effects as in the motivated model: (1) No Trend, (2) Smooth Trend, (3) Time Fixed Effects.

The regions are set as follows: The first regions are $A_1 = (0, 20]$ and $B_1 = (0, 200]$, and correspond to the short age group and a group with very small number of products in the same category. The remaining regions $A_2, ..., A_6$ are determined as intervals with the endpoints defined by the quantiles of Age $T_{i,t}$ with indices

$$.2, .4, .6, .7, .8, 1.$$

The remaining regions $B_2, ..., B_6$ are determined as intervals defined by the quantiles of the Number of Products by Category $N_{i,t}$ with indices $.2, .4, .6, .7, .8, 1$, as well as values $200, 1000, 2000$. The resulting regions are reported as part of the regression results in Table 3.

The same remarks as before apply here regarding the inclusion of fixed effects. Note that in this specification the base prediction is the one using the small number of products, $N < 200$, or short age, $T < 20$. The coefficients measure how our prediction changes as we go through the various regions for age or for the number of products in the same category.

3.2. **Empirical Results for the Motivated Model.** Table 2 shows the results for the theory-motivated models with product fixed effects, with three specifications for the time effects: (1) No Trend, (2) Smooth Trend, and (3) Time Fixed Effects. We

report there the estimated coefficients as well as the (standard) standard errors that have been clustered by product and time.[4]

The signs on the coefficients for the terms containing age T are consistent with the model, and the signs for the N follow less systematic patterns. We examine the T and N effects graphically. We show in Figures 6, 7, and 9 the estimated Product Age (T) Effect, the Number of Products (N) Effect, and the Time Effects on the predicted probability of Big Forecast Error Event:

R1: The T effect is the strongest in the model without trends, reaching $-.35$ for high T. The T effect is more modest in the models that allow for smooth trends or time effects, flattening at $-.1$ in the time effects models. There are essentially no improvements due to large T in the two models that adjust for time effects. As we argue below, the true effect trajectory probably lies in a convex combination of these reported effect trajectories. Overall, the T effect seems to agree with predictions of the theoretical model, and there are diminishing returns (improvement in forecast error) to large T.

R2: Without time adjustments, the N effect seems to be strong early on, but it does exhibit diminishing returns to scale and saturates at $-.4$ once $N > 5000$. Note, however, that this effect is *not statistically significant*, and we can't reject the hypothesis of a flat effect with respect to N, see Table 2. If we inspect the model with the smooth trends, we see that the N effect takes values of about -.25 at moderate N levels (relative to the inception point) and then attenuates toward zero at high N. This effect is also *not statistically significant*. [5] This contradicts the predictions of the theoretical model.

[4]The reported standard errors probably only partially account for the dependence in the data, which is quite complex due to the presence of forecast learning, entering the definition of the dependent variable through $\hat{Q}_{i,t}$. We followed a good empirical practice here, and note that our problem is about getting point estimates of the effect of data size on quality of forecast learning, and not about the (precise) significance testing.

[5]In the third model with unrestricted time effects, large N seems to moderately increase the probability of the big error event.

This modestly harmful effect of increasing N does appear to be significant, at least given the conventional time-product clustered standard errors we use. In summary, the N effect seems to exhibit either positive, quickly diminishing returns to scale or modestly negative dis-returns, depending on how we control for time effects.

R3: The estimated time effects indicate that there is a general improvement in the relative forecast error over time; this could reflect improvements in the forecasting engine itself, for example using more features (traffic data etc.), as well as the time evolution of the mix of products.

We need to interpret the empirical result R1 in light of R3. The ages $T_{i,t}$ and trends t are correlated in this data. Hence the model without trends attributes the aggregate changes in the occurrence of the BigErrorEvent, shown in Figure 3, to the longer T effect, whereas the model with the trends or time effects projects out the aggregate changes or trends, before it attributes the effect to longer T. The aggregate changes, shown in Figure 3, suggest gradual improvements in forecasting performance, which occur due to *both*

- (1) methodological improvements (e.g. using more features), and
- (2) availability of longer histories/ages (T) for many products.

Hence the model without trends attributes the aggregate changes to source (2). The Time Effects model attributes the aggregate changes captured by time effects to source (1) and the remainder of the effects for source (2). The Smooth Trend model attributes aggregate changes captured by a smooth trend to source (1) and the rest of the changes to source (2).

3.3. **Empirical Results for the Agnostic Model.** In Table 3 we show the results for the agnostic model with three specifications for the time effects: (1) No Trend, (2)

TABLE 2. Estimation Results of Fixed Effects Models for the Motivated Model

	Dependent variable:		
	1(Relative Forect Error $> X$)		
	Without Trend	Time Trend	Time Effects
	(1)	(2)	(3)
Age > 20	-0.800^{***}	-0.029	-0.140
	(0.046)	(0.086)	(0.085)
(Age>20) Inv Root Age	9.230^{***}	0.395	1.460
	(0.574)	(0.948)	(0.936)
(Age> 20) Inverse Age	-27.472^{***}	-1.717	-4.247
	(1.826)	(2.678)	(2.643)
N> 200	-0.487^{**}	0.053	0.292
	(0.212)	(0.229)	(0.237)
(N>200) Inverse Root N	8.534^{*}	-6.234	-13.931^{**}
	(4.608)	(5.106)	(5.422)
(N> 200) Inverse N	-23.457	81.345	153.613^{**}
	(62.620)	(65.474)	(67.159)
Trend		-0.327^{***}	
		(0.075)	
Squared.Trend		-0.060	
		(0.045)	
Observations	496,259	496,259	496,259
R^2	0.276	0.278	0.284
Adjusted R^2	0.268	0.270	0.277

Note: Standard errors are clustered by product and date.

Smooth Trend, and (3) Time Fixed Effects. We report the estimated coefficients as well as the standard errors that have been clustered by product and time.[6]

We show in Figures 10, 11, and 12 the estimated Product Age (T) Effect, the Number of Products (N) Effect, and the Time Effects on the predicted probability of Big Forecast Error Event:

AR1: The early T effect is the strongest in the model without trends, reaching -.5 to -.6 for the high T group. The T effect is modest in models that allow for smooth trends or time effects, reaching the magnitude of -.05 to -.07 in the high T group. The true effect is probably between these two reported effects, by the argument we gave for the result R2.

AR2: The N effect seems flat, sometimes even negative, and is qualitatively similar across three different specifications. Note that here the point estimates from the Agnostic Model deviate from those of the Motivated Model without time adjustments, but the differences are not statistically significant, despite the visual differences in plotted effects.

AR3: The estimated Smooth Trend and Time Effects indicate that there is a general improvement in the relative forecast error over time; this could reflect improvements in the forecasting model itself or, as discussed previously, could also capture improvements due to the average histories getting longer.

Overall Conclusions. In both empirical specifications, age T, and, to a much smaller extent, number of products N, explain a portion of the forecast accuracy. Data suggests there are diminishing returns to T, and the effects tend to reach saturation once T reaches medium sizes in some models. The early effect of T could be substantial or small, depending on how we attribute the overall trends in forecasting errors. Forecast errors overall show general but gradual improvement in time.

[6]The previous comment about standard errors applies here as well.

TABLE 3. Estimation Results of Fixed Effects Models for the Agnostic Model

	Dependent variable:		
	1(Relative Forect Error $> X$)		
	Without Trend	Time Trend	Time Effects
	(1)	(2)	(3)
Age.Region.(20,58]	-0.059^{***}	-0.024^{***}	-0.017^{**}
	(0.006)	(0.007)	(0.007)
Age.Region.(58,99]	-0.098^{***}	-0.018^{*}	-0.019^{**}
	(0.008)	(0.010)	(0.009)
Age.Region.(99,148]	-0.167^{***}	-0.033^{***}	-0.031^{***}
	(0.010)	(0.013)	(0.012)
Age.Region.(148,199]	-0.265^{***}	-0.071^{***}	-0.047^{***}
	(0.012)	(0.017)	(0.015)
Age.Region.(199,278]	-0.342^{***}	-0.076^{***}	-0.047^{***}
	(0.014)	(0.021)	(0.017)
N.Region.(200,1e+03]	-0.028	-0.007	0.020
	(0.061)	(0.061)	(0.060)
N.Region.(2e+03,3.72e+03]	0.055	0.071^{*}	0.014
	(0.036)	(0.037)	(0.036)
N.Region.(3.72e+03,7.74e+03]	0.050	0.088^{**}	0.021
	(0.038)	(0.039)	(0.038)
N.Region.(7.74e+03,9.65e+03]	0.059	0.106^{***}	0.031
	(0.040)	(0.041)	(0.040)
N.Region.(9.65e+03,1.07e+04]	0.035	0.099^{**}	0.027
	(0.040)	(0.041)	(0.040)
N.Region.(1.07e+04,1.57e+04]	0.064	0.134^{***}	0.075^{*}
	(0.041)	(0.042)	(0.041)
Trend		-0.262^{***}	
		(0.063)	
Squared.Trend		-0.056	
		(0.044)	
Observations	496,275	496,275	496,275
R^2	0.276	0.278	0.284
Adjusted R^2	0.269	0.271	0.277

Note: Standard errors are clustered by product and date.

Data suggest that there do not seem to be statistically significant positive returns to the number of products N, and the estimated effect of N is actually negative (and statistically significant in at least two models). The latter finding does not accord well with the stylized theoretical model, though somewhat aligned with predictions from the "extended" theoretical model in Section 2.3, with the "slow" effect N^{-a} with $a \approx 0$.

4. Results for All Other Major Product Groups: Summary

The purpose of this section is to generalize the analysis to all 36 major product groups. [7] We define the relative forecast error as before, namely as

$$|Q_{i,t} - \hat{Q}_{i,t}|/(Q_{i,t} + 1),$$

and the following dependent variable

$$Y_{i,t} = 1\{\underbrace{|\hat{Q}_{i,t} - Q_{i,t}|/(Q_{i,t} + 1) > X}_{\text{"Big Forecast Error Event"}}\}.$$

This variable describes the event of making a *big forecast error*, namely $X\%$ error. The threshold of X is chosen so that, within given product group, this threshold is exceeded for any random product-week pair (i, t) with probability $P = 30\%$. We maintain the same *uniform rule* for $P = 30\%$ for *all* 36 product groups, to determine the thresholds X. A nice feature of this uniform-in-product group "P rule" is that it allows us, on the one hand, to treat all products group equally, avoiding arbitrariness, and, on the other hand, allows for automatic adjustment of threshold X for the obvious heterogeneity across products groups (some product groups are much harder to forecast than others).

[7] These product groups include Apparel; Automotive; Baby; Beauty; Business, Industrial & Scientific Supplies; Books; Camera; Electronics; Furniture; Grocery; Health & Personal Care; Home; Home Entertainment; Home Improvement; Jewerly; Kitchen; Lawn and Garden; Luggage; Luxury Beauty; Major Appliances; Music; Musical Instruments; Office Products; Outdoors; PC; Pantry; Pet Products; Shoes; Software; Sports; Tools; Toys; Video & DVD; Video Games; Watches; and Wireless.

We would like to estimate a predictive model for the effect of $N_{i,t}$ and $T_{i,t}$ on the probability of the Big Error Event. We shall employ the Agnostic Model, where we divide the ranges of $T_{i,t}$ and $N_{i,t}$ into regions, as well as product and time effects:

$$Y_{i,t} = \underbrace{\alpha_i + \beta_t}_{\text{Product and Time Effects}} + \underbrace{\sum_{j=1}^{5} \delta_j 1(T_{i,t} \in A_j)}_{\text{Product Age Effect}} + \underbrace{\sum_{k=1}^{5} \gamma_k 1(N_{i,t} \in B_k)}_{\text{Number of Products Effect}} + \epsilon_{i,t},$$

where the projection error $\epsilon_{i,t}$ is by definition taken to be orthogonal to the space spanned by all variables, including the product fixed effects and time effects. As before, the terms α_i are the product specific fixed effects, which capture the notion that some products are inherently harder to forecast than others. The terms β_t are the time effects, which capture that on some dates it may be harder or easier to forecast than on others; moreover, these time effects also partly capture the overall improvement in the forecasting model that produces $\hat{Q}_{i,t}$. As before, we consider three specifications for time effects:

(1) No Trend: $\beta_t = 0$ for all $t = 1, ..., H$.
(2) Smooth Trend: $\beta_t = b_1(t/H) + b_2(t/H)^2$, where for some parameters b_1 and b_2,
(3) Time Fixed Effects: β_t's are unrestricted parameters.

The regions are set as follows: The first regions are $A_1 = (0, 20]$ and $B_1 = (0, 200]$, and correspond to the short age group and a group with very small number of products in the same category. The remaining regions $A_2, ..., A_6$ are determined as intervals with the endpoints defined by the quantiles of Age $T_{i,t}$, *within the given product group*, with indices

$$.2, .4, .6, .7, .8, 1.$$

The remaining regions $B_2, ..., B_6$ are determined as intervals defined by the quantiles of the Number of Products by Category $N_{i,t}$ with indices $.2, .4, .6, .7, .8, 1$, *within*

the given product group, as well as values $200, 1000, 2000$. Not all regions will be populated for all products groups, in which case the estimation results will return no estimate of the coefficients corresponding to that region.

We are focusing our analysis on the agnostic model, since it is a flexible, easy-to-interpret model and allows the data to speak for itself. The model is still informed by the theory in Section 2 in that, as before, we are linking the probability of the big error event to the data sizes $T_{i,t}$ and $N_{i,t}$ in a general manner prescribed by the theory, but we relax the functional form so that the empirical results can easily differ from the predictions derived from the stylized form of the theory. By doing so, for example, we are allowing the model to capture the N and T effects of the form allowed for in the "extended" theoretical model.

We will summarize the results graphically, but a full set of empirical results recorded in a table form is available on request. We show in Figures 13-15, and Figures 16-18 the estimated Product Age (T) Effect and the Number of Products (N) Effect for all other major product groups. Figure 19 shows the estimated Time Effects for the agnostic model with the time effects for all major product groups. Overall, we find that the results qualitatively agree with those for the case of Electronics.

GR1: The T effect is the strongest in the model without trends, reaching -.3 or -.6 in the high T bucket for some product groups – for instance Apparel, Shoes, Watches and Wireless. The T effect is more modest in models that allow smooth trends or time effects. For many products, there are essentially no detectable improvements due to large T in the two models that adjust for time effects. Some more notable exceptions include Apparel, Shoes, and Video Games product groups, where the effect bottoms out at about -.2 and -.25. As we argue before in the case of Electronics, the true effect probably lies in a convex combination of these reported effects. Overall, once we control for time effects, the effect of age T exhibits small, diminishing

returns to scale. Without controlling for time, increases in T can keep substantially improving the forecast quality in some cases that we noted above.

GR2: The N effect appears to be qualitatively similar across three specifications for time effects, so we will focus on the most flexible specification with the unrestricted time effects. The N effect seems essentially flat for many product groups, with a few exceptions. For example, for Apparel, Kitchen, PC, and Shoes product groups, the N effect reaches $-.3$ to $-.5$ early on, but stays essentially flat at that level, thereby exhibiting diminishing returns to scale. However, for example, for Home product groups, the estimated N effect is actually associated with higher relative forecast errors, reaching a large positive magnitude for early buckets of N, and then staying flat at that level. By inspecting the figures we can classify about 90% of other product groups into cases where the N effect is associated with either a $.15$ decrease or increase in the probability of the big error event. In summary, the N effect, when not already being essentially flat, seems to exhibit either positive (or negative), quickly diminishing returns to scale.

GR3: The estimated time effects indicate that there is a general improvement in the relative forecast error over time; this could reflect improvements in the forecasting engine itself, for example, due to using more features (traffic data etc.), as well as the time evolution of the mix of products.

As in the case of Electronics, we need to interpret the empirical result GR1 in light of GR3. The ages $T_{i,t}$ and trends t are highly correlated in this data. Hence the model without trends attributes aggregate changes in the occurrence of the BigErrorEvent to the longer T effect, whereas the models with the trends or time effects project the aggregate changes or trends out, before they do the attribution to the longer T effect. As we argued before, the true effect of age is thus likely to be bounded between these two effects.

5. Conclusion

We have developed theoretical and empirical evidence on the impact of the size of panel data sets on the quality of demand forecasting. Empirical evidence utilizes a large data set on all major product groups from Amazon.com. Theoretical evidence makes use of a state-of-the art theoretical learning model in order to see what can happen *in principle* to the quality of forecast as the data size becomes larger. The data size dimensions include N, the number of products in a broad similar category, and T, the length of available histories for a given product. In the theoretical learning model, there are diminishing returns to the size of data, and the returns become flatter in non-stationary environments. Our theoretical model also informs our empirical analysis, where we consider a "theory motivated" model and an "agnostic" model, where the first model allows for some deviations from the functional form prescribed by the theoretical model, and the second model allows for more flexibility, to let data speak for itself.

Generally we find that T, the length of histories, is robustly helpful in improving the forecast quality. The effect of N, the number of products in the same category, is robustly flat (with a few exceptions). When the estimated effects of T and N are not flat, they exhibit diminishing returns to scale, with the exception of T effects in the model without time controls.

We believe that our results are helpful in terms of understanding how data size influences firm performance. In particular, asymptotic/statistical learning theory typically tells us that data in general has diminishing returns. The accuracy of our forecasting models improves at a decreasing rate as the sample size increases. While this result is unsurprising from the viewpoint of asymptotic theory, the connection to the broader policy issues seems not to have been widely made. In particular, we do not see evidence for a version of the "data feedback loop" theory, wherein adding new products leads to an indirect network effect with more accurate forecasts leading to more customers/sales leading in turn to more accurate

forecasts. As we demonstrate, velocity does not improve the accuracy of percentage forecast errors and indeed decreases forecast accuracy in levels. We do not find that adding new products add to forecast accuracy, except perhaps in the case where the number of products in a product line is very small.

We find instead that the performance of our forecasting models is improving in product age. This, again, is different than a naive description of a data feedback loop based on indirect network effects caused by the addition of new products. The source of improvement has a very simple and fundamental root cause: the longer the time series, the better we can learn the parameters of the demand forecast, be it a standard time series forecasting model or the state-of-the art augmented factor model used in out theoretical benchmarking. Also, models seem to demonstrate a trend level of improvement which is most likely associated with the trial and error of the scientific method and investment in better engineering and infrastructure.

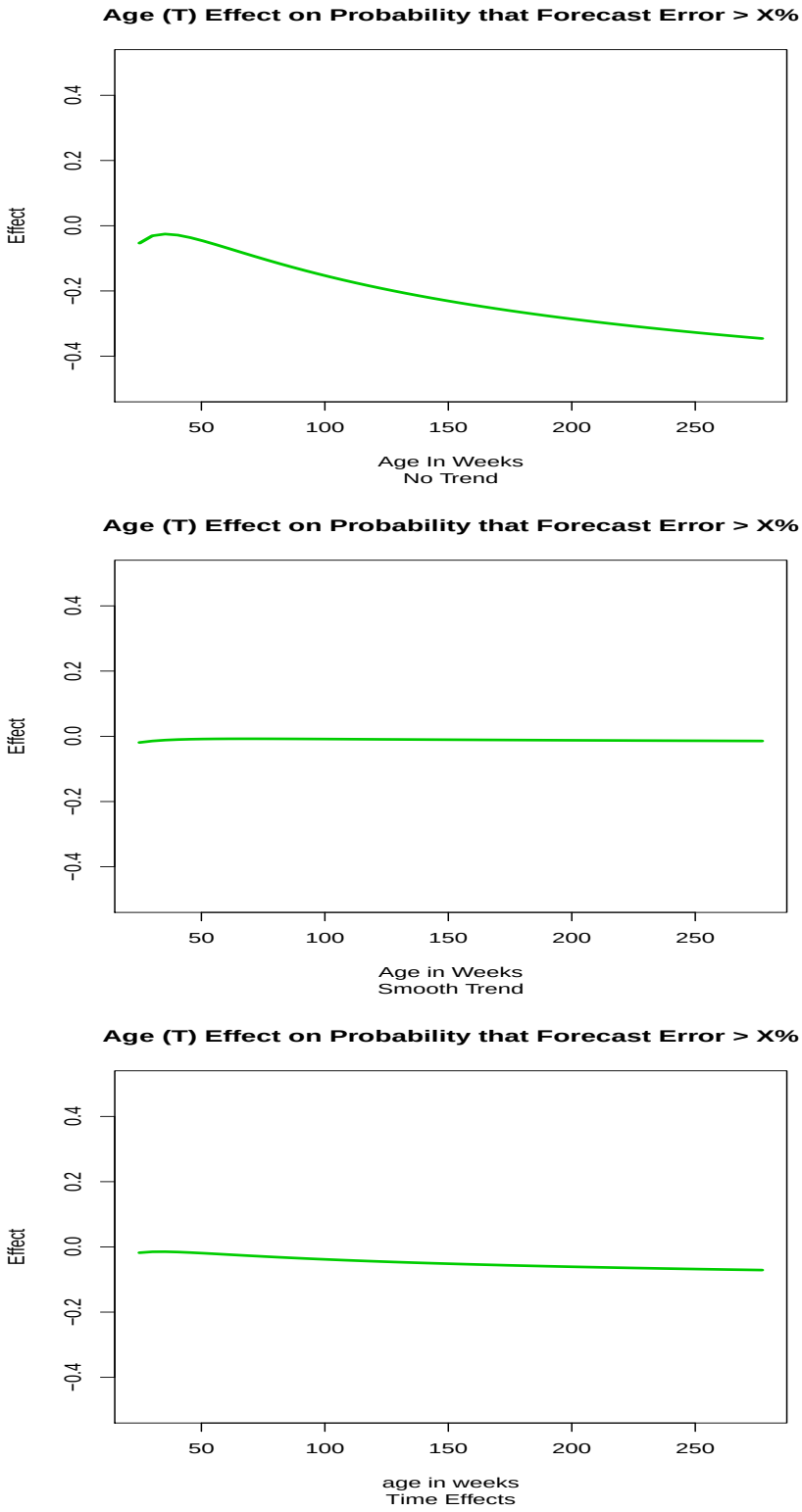


FIGURE 6. The Estimated Impact of T on the Quality of Learning in the Motivated Model with No Trend, Smooth Trend, and Time Effects

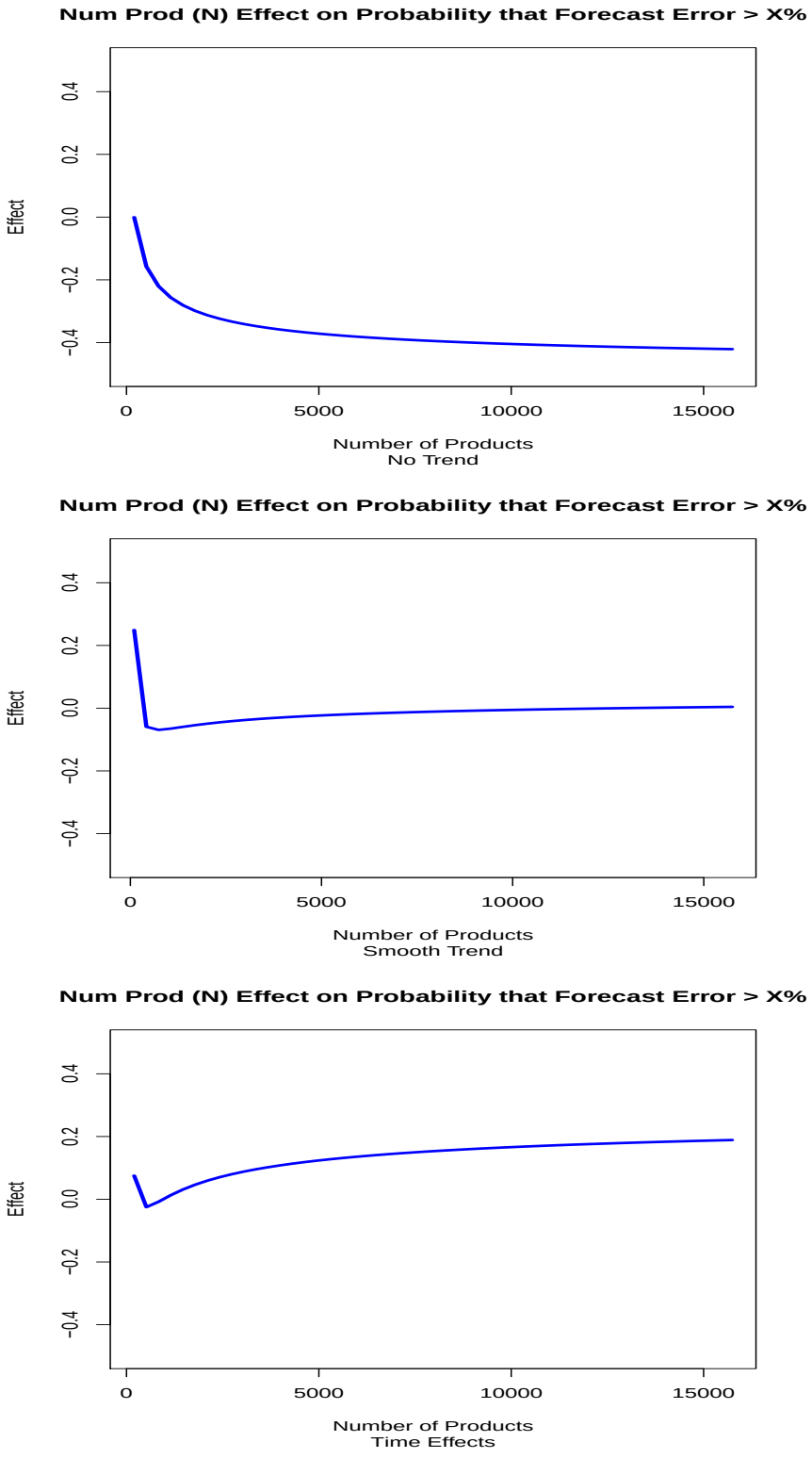


FIGURE 7. The Estimated Impact of N on the Quality of Learning in the Motivated Model with No Trend, Smooth Trend, and Time Effects

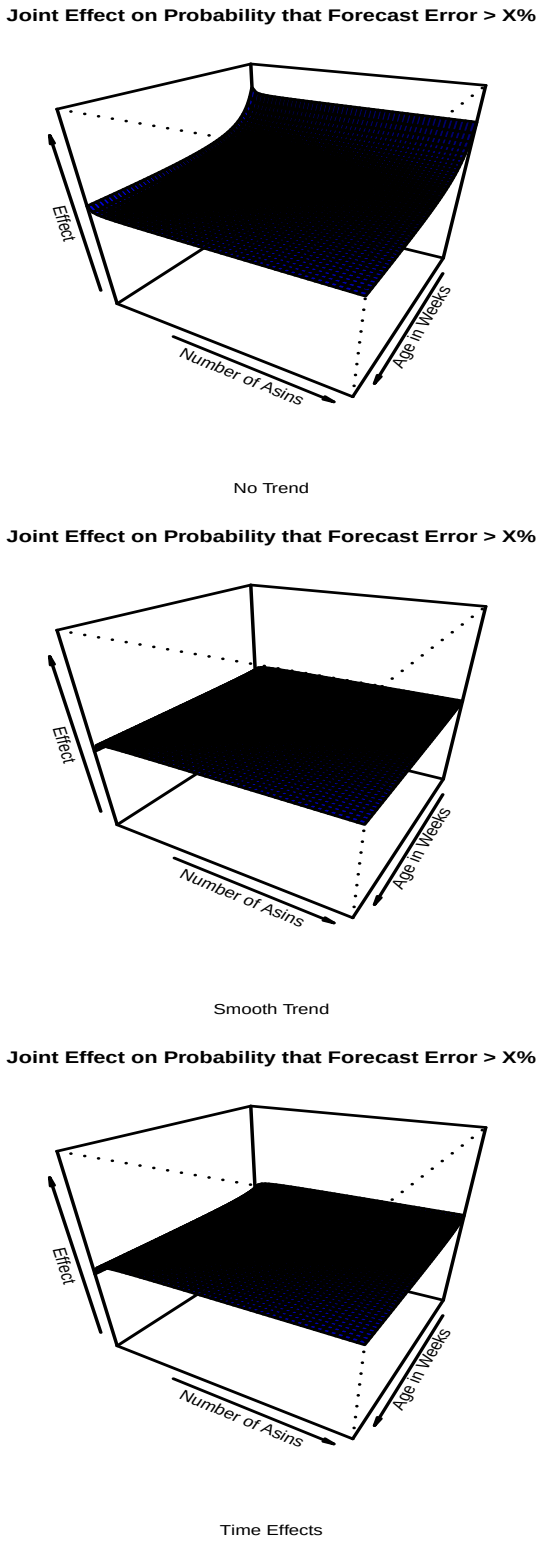


FIGURE 8. The Estimated Impact of N and T on the Quality of Learning in the Motivated Model with No Trend, Smooth Trend, and Time Effects

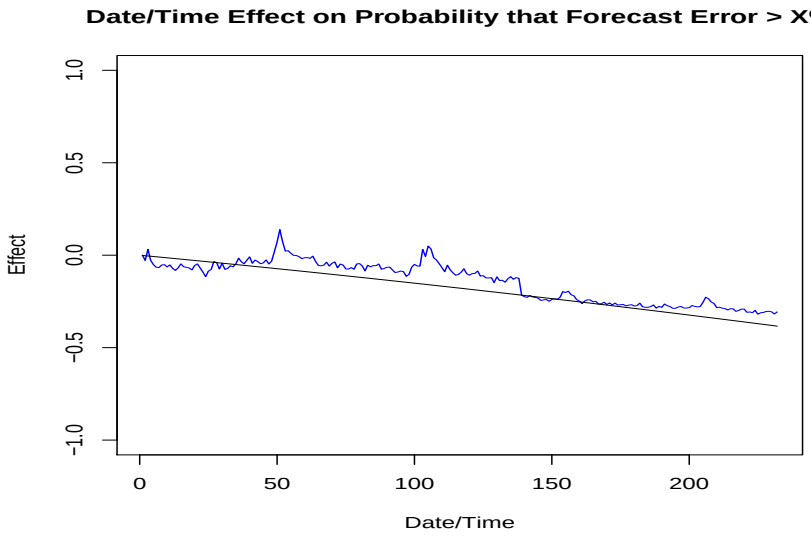
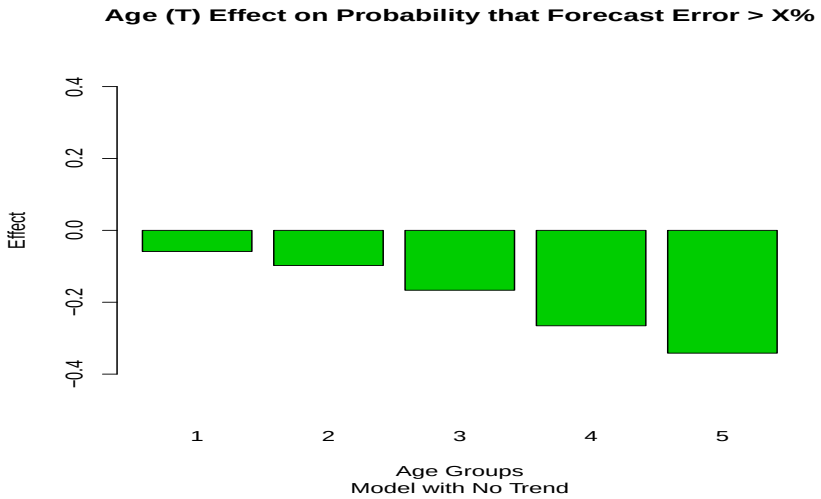
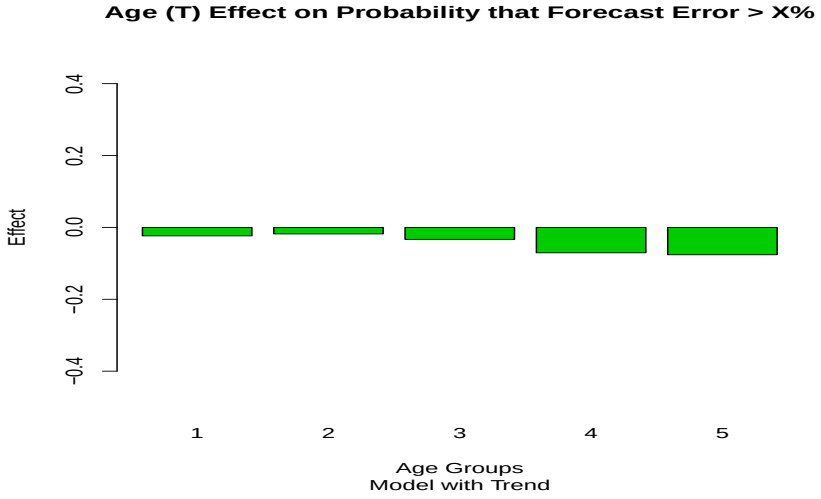


FIGURE 9. The Estimated Date/Time Effects on the Quality of Learning in the Motivated Model with Smooth Trend and Time Effects





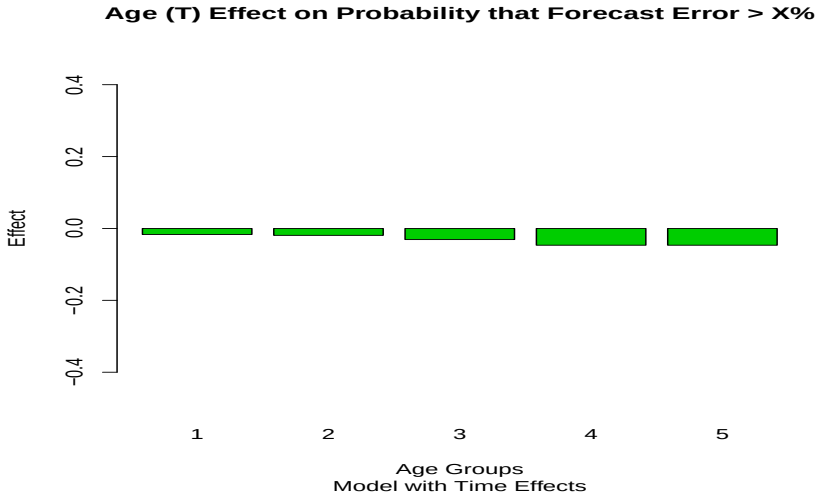
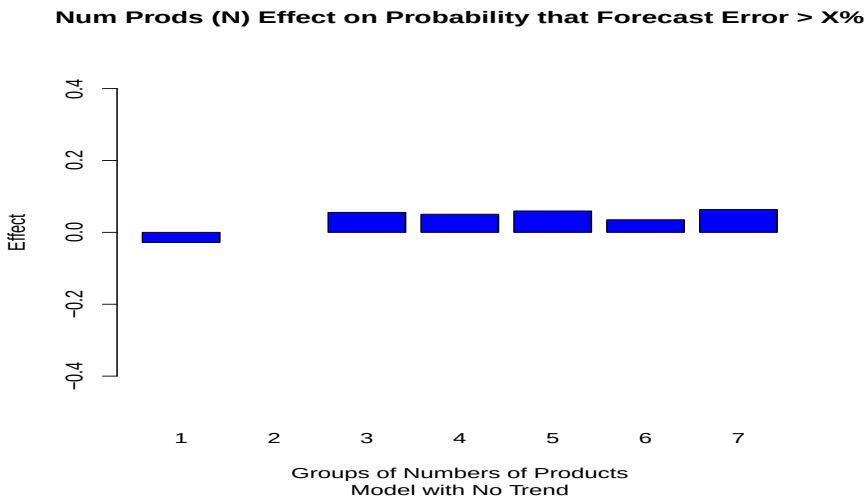
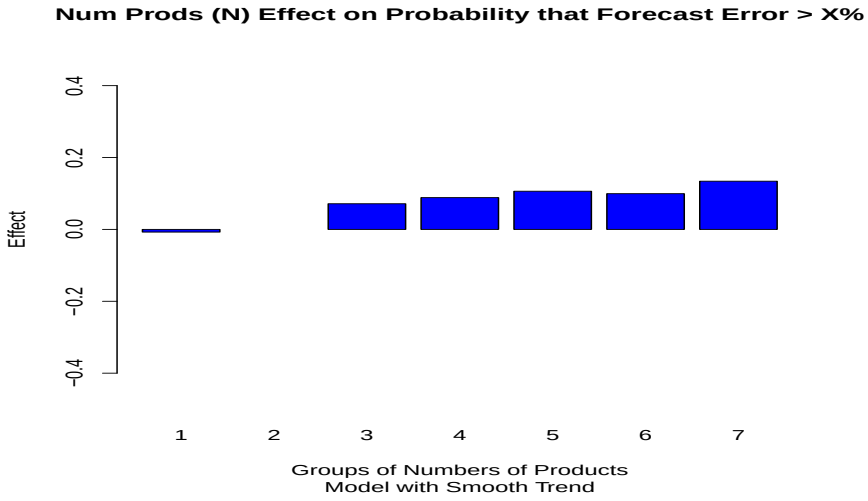


FIGURE 10. The Estimated Impact of T on the Quality of Learning in the Agnostic Model with No Trend, Smooth Trend, and Time Effects





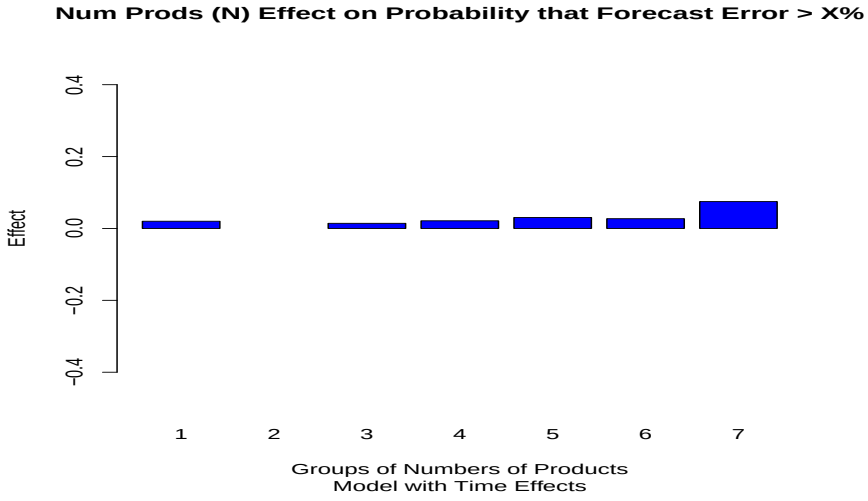


FIGURE 11. The Estimated Impact of N on the Quality of Learning in the Agnostic Model with No Trend, Smooth Trend, and Time Effects

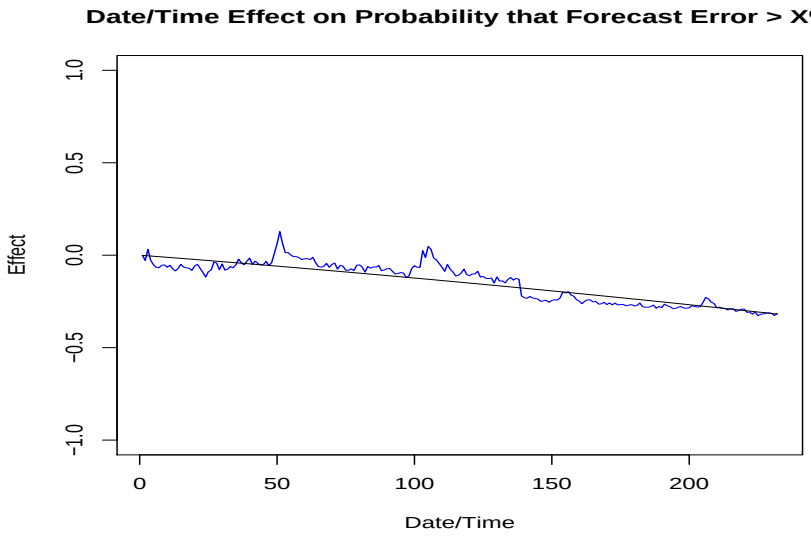


FIGURE 12. The Estimated Date/Time Effects on the Quality of Learning in the Agnostic Model with Smooth Trend and Time Effects

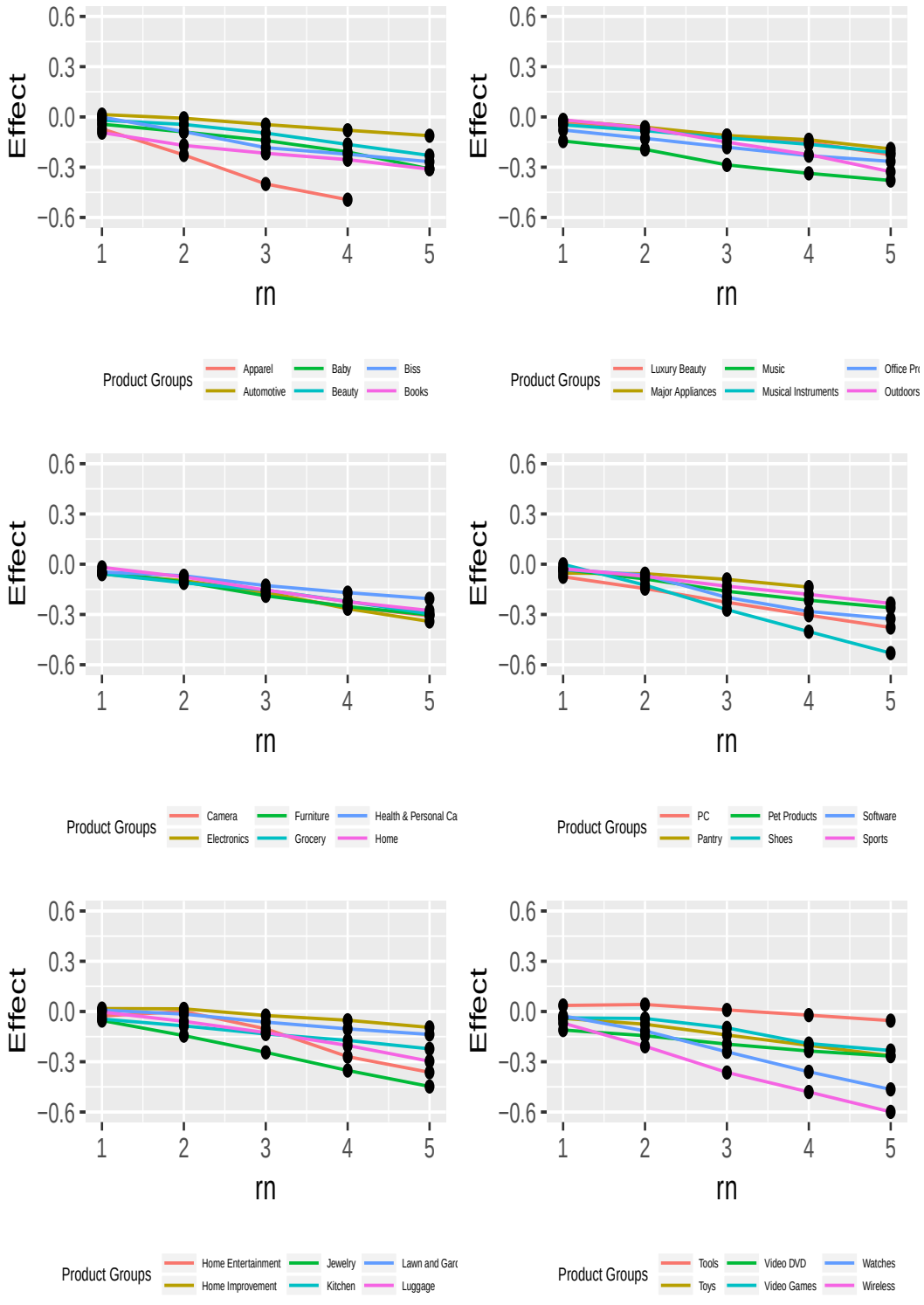


FIGURE 13. The Estimated Impact of T on the Quality of Learning in the Agnostic Model with No Trends, applied to all product groups. Note that the label "rn" stands for the age group.

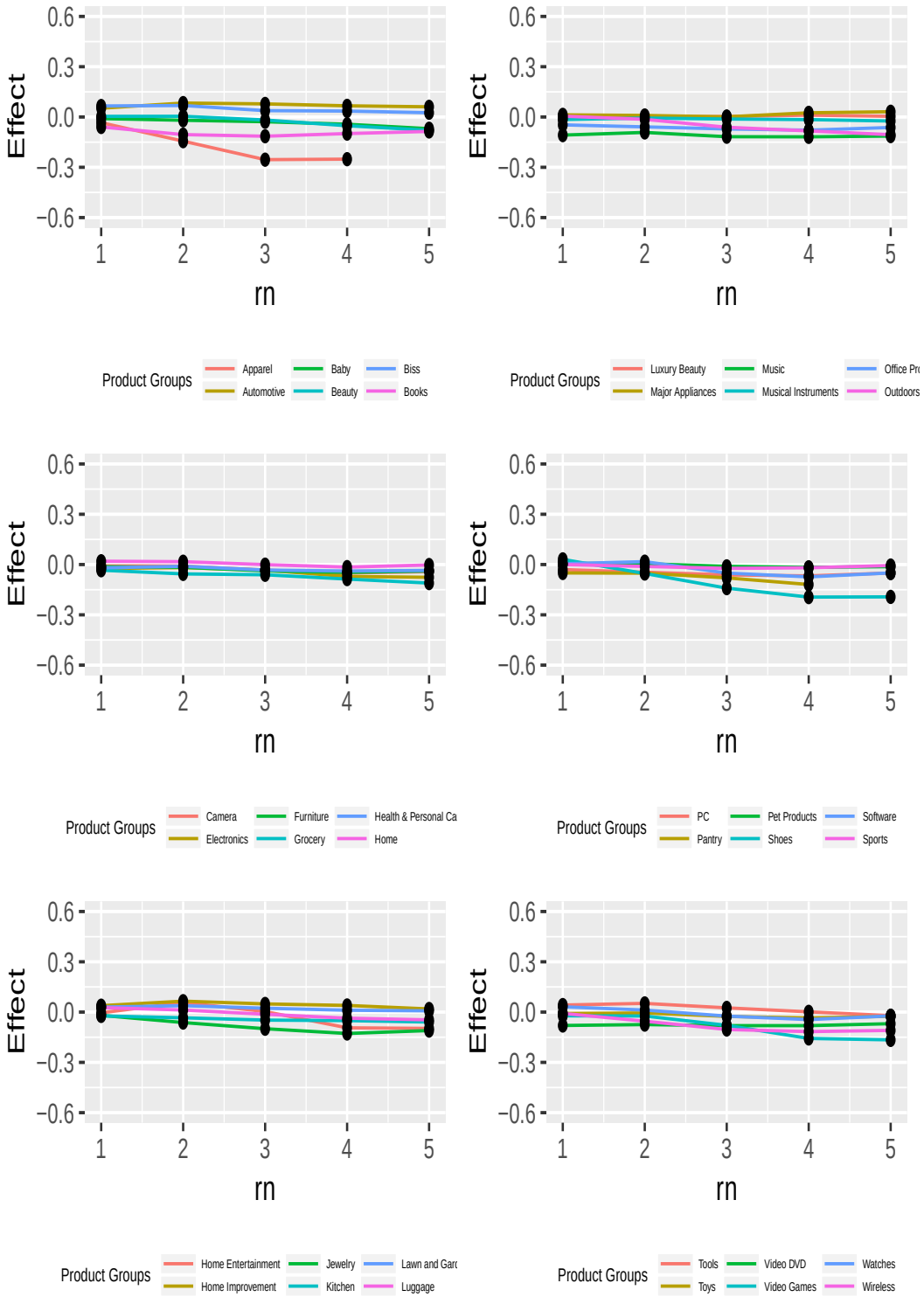


FIGURE 14. The Estimated Impact of T on the Quality of Learning in the Agnostic Model with Smooth Trends, applied to all all product groups. Note that the label "rn" stands for the age group.

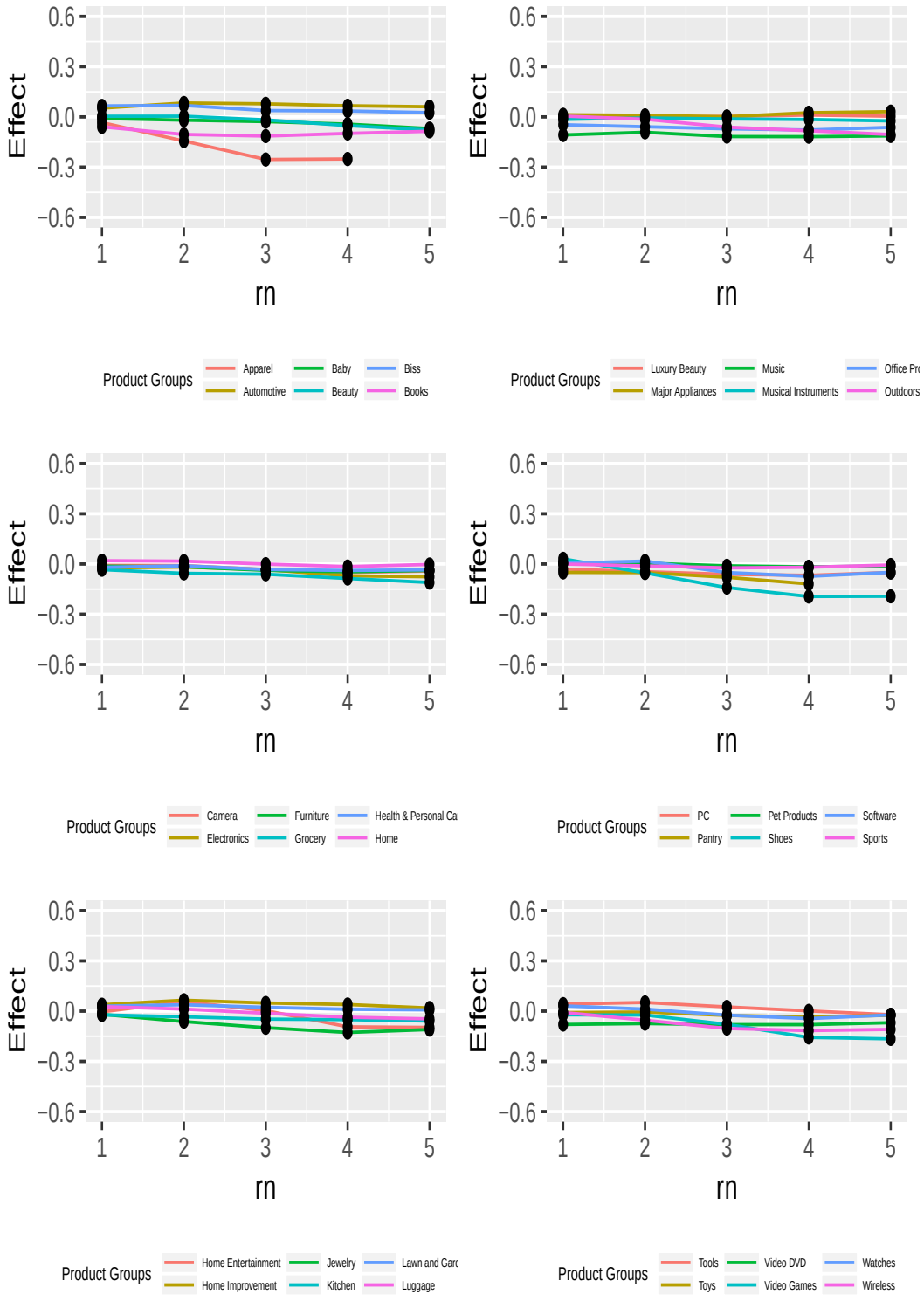


FIGURE 15. The Estimated Impact of T on the Quality of Learning in the Agnostic Model with Time Effects, applied to all all product groups. Note that the label "rn" stands for the age group.

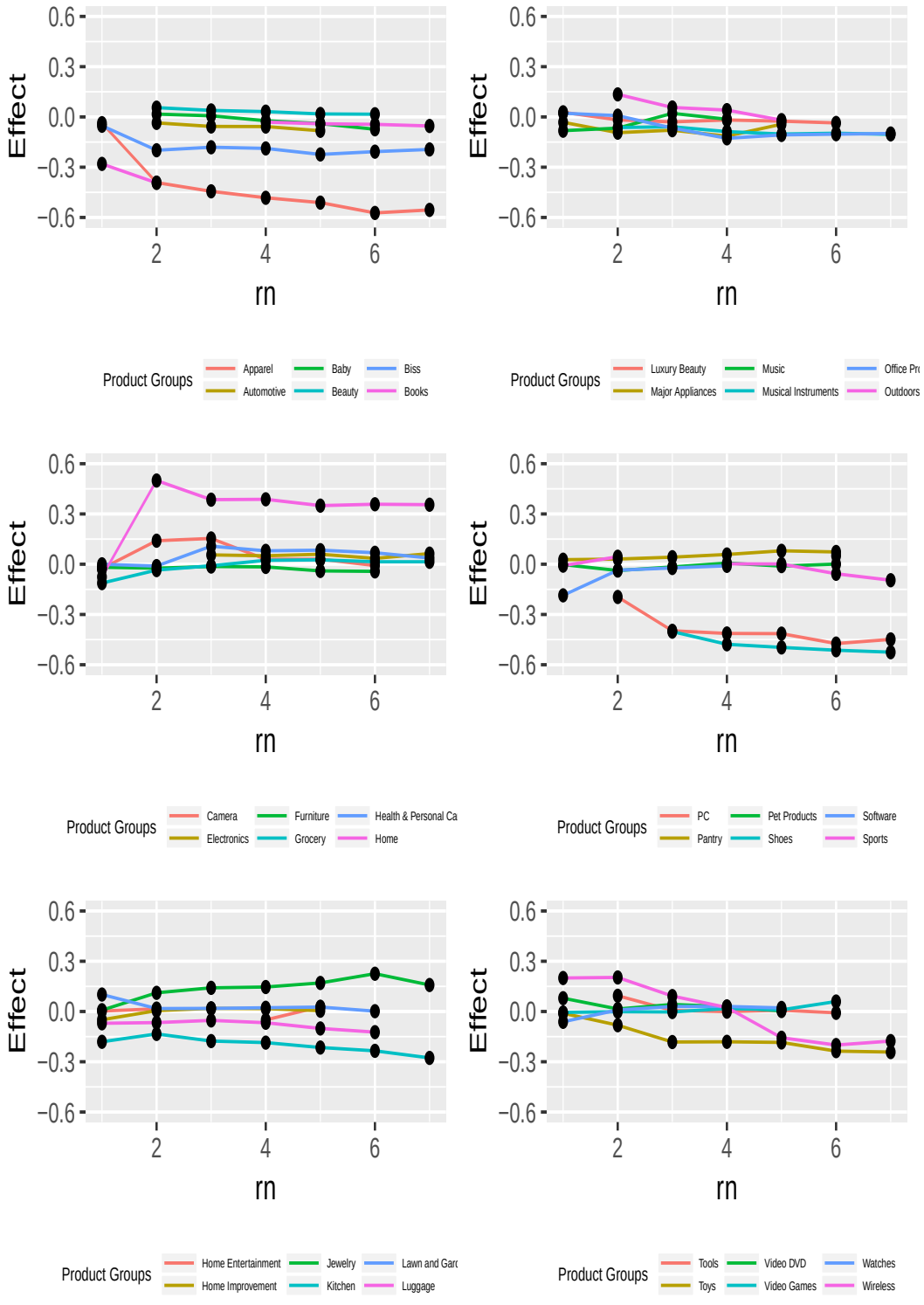


FIGURE 16. The Estimated Impact of N on the Quality of Learning in the Agnostic Model with No Trends, applied to all product groups. Note that the label "rn" stands for the N group.

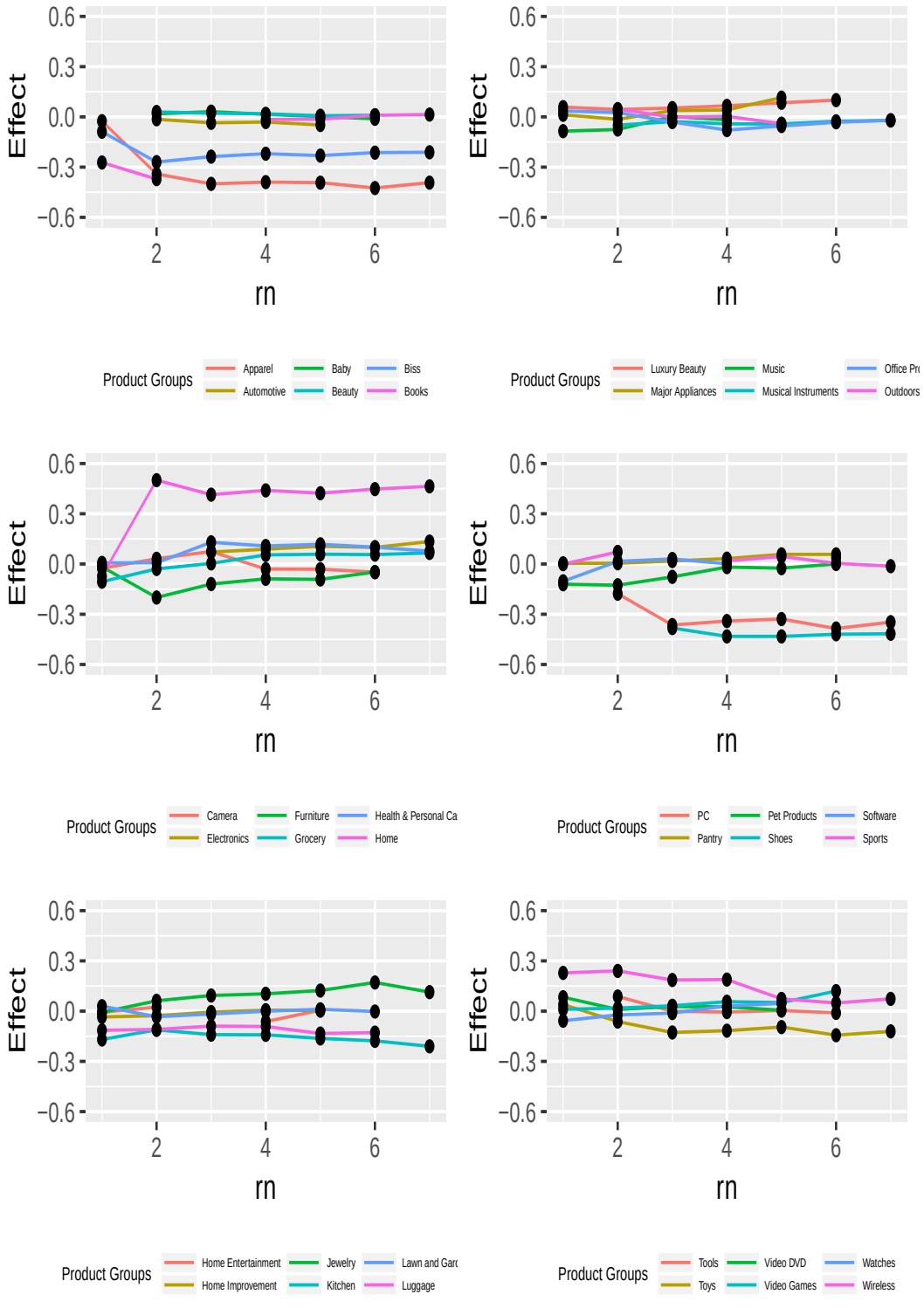


FIGURE 17. The Estimated Impact of N on the Quality of Learning in the Agnostic Model with Smooth Trends, applied to all product groups. Note that the label "rn" stands for the N group

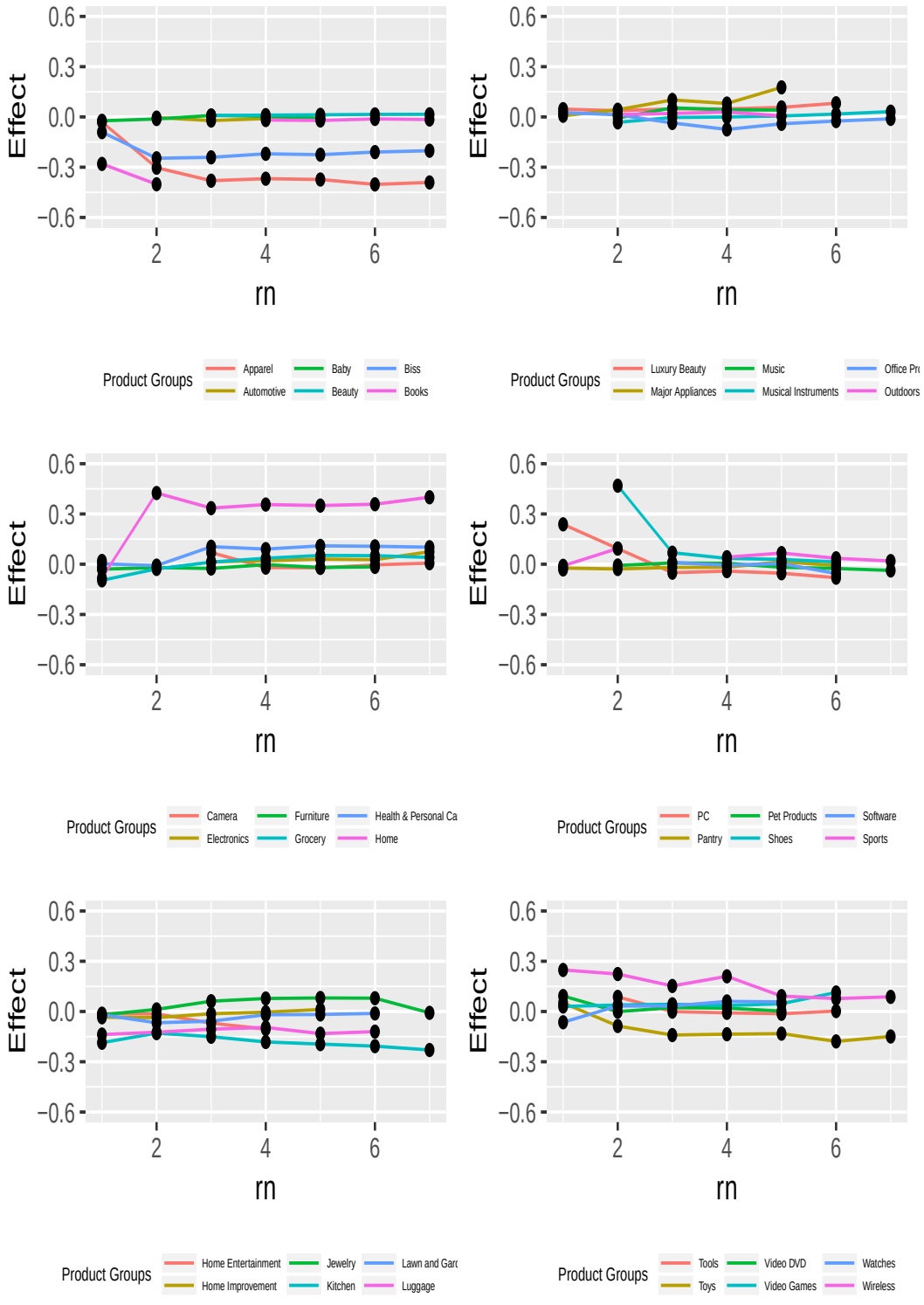


FIGURE 18. The Estimated Impact of N on the Quality of Learning in the Agnostic Model with Time Effects, applied to all product groups. Note that the label "rn" stands for the N group.

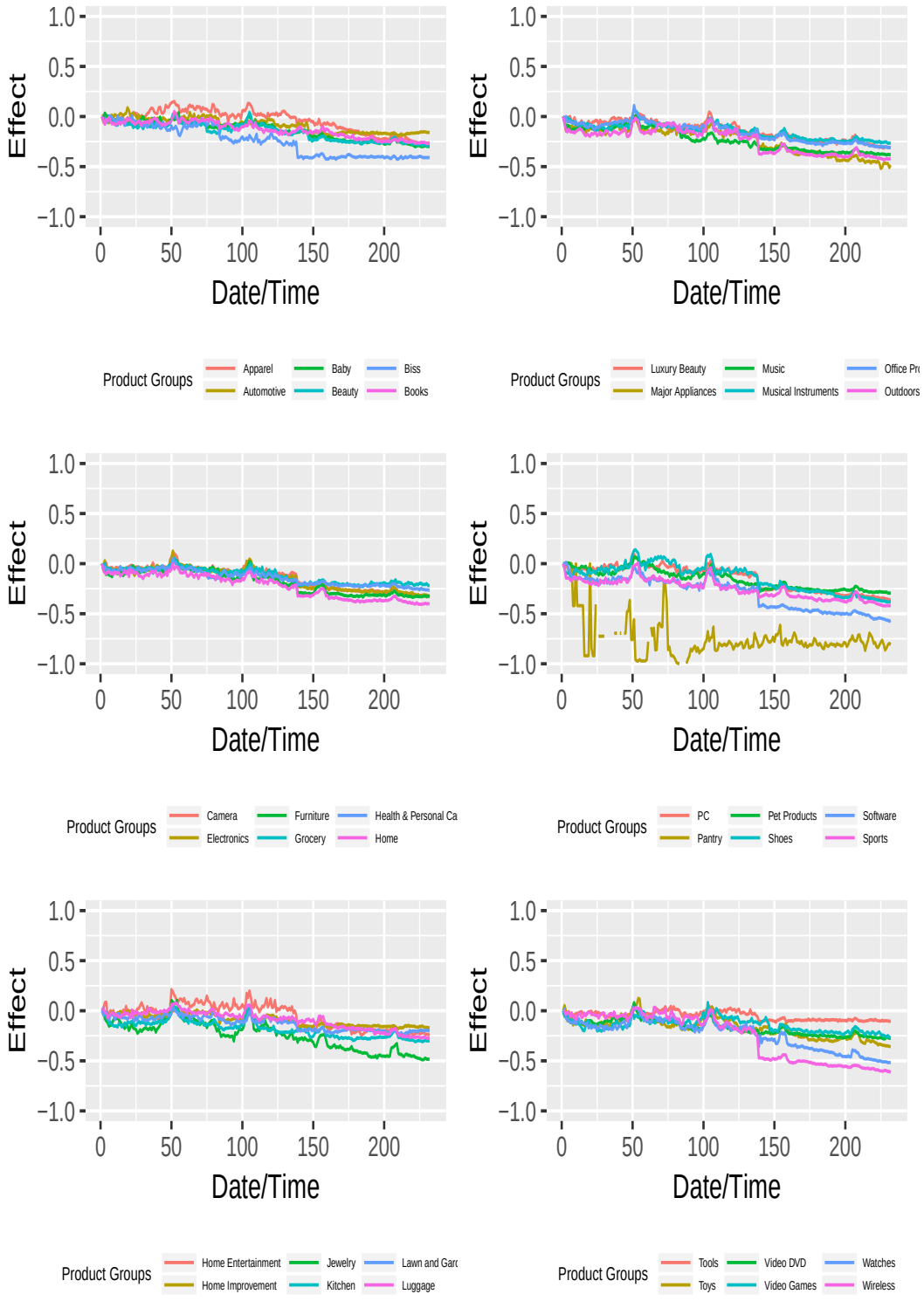


FIGURE 19. The Estimated Time Effects in the Agnostic Model with Time Effects, applied to all product groups.

References

Bai, Jushan. "Panel Data Models with Interactive Fixed Effects." *Econometrica*, 2009, 77(4), 1229-352.

Bai, Jushan, and Serena Ng. "Determining the number of factors in approximate factor models." *Econometrica* 2002, 70(1), 191-221.

Bernanke, Ben S., Jean Boivin, and Piotr Eliasz, "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach", NBER working paper no. 10220, January 2004.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *Quarterly Journal of Economics*, 2002, 117(1), 339-76.

Bloom, Nicholas R., Rafaella Sadun, and John van Reenen, "American do IT better: U.S. Multinationals and the productivity miracle," *American Economic Review*, 2012, 102(1), 167-201.

Chandler, A. D., *The Visible Hand: The Managerial Revolution in American Business*, 1977, Cambridge, MA: The Belknap Press.

de Fortuny, Enric J., David Martens, and Foster Provost, "Predictive Modeling with Big Data: Is Bigger Really Better?", *Big Data*, 2013, 1(4), 215-26.

Grunes, Allen P. and Maurice E. Stucke, "No Mistake about it: The Important Role of Antitrust in the Era of Big Data," 2015, University of Tennessee Legal Studies Research Paper No. 269.

Lambrecht, Anja and Catherine Tucker, "Can Big Data Protect a Firm From Competition," 2017, *Competition Policy International Antitrust Chronicle*, January 2017, 1-8.

Lerner, Andreas, "The Role of 'Big Data' in Online Platform Competition," 2014, working paper.

McElheran, Kristina, and W. Jin, "Economies before scale, I.T. investment and performance in young firms," working paper, University of Toronto.

Newman, Nathan, "Search, Antitrust and the Economics of the Control of User Data," 2014, working paper, NYU Information Law Institute.

Tambe, Prasanna and Lorin Hitt, "The productivity of Information Technology Investments: New Evidence from IT labor data," *Information Systems Research*, 2012, 23(3-part-1), 599-617.

Varian, Hal, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*," 2014, 28(2), 3-28.

Appendix A. Derivation of Assertion 2

Here we first explain the estimation of a_i via the plug-in principle. The plug-in estimator is defined by:

$$\hat{a}_i = (\hat{\sigma}_i^2 + \hat{\alpha}_i' \hat{\Sigma}_v \hat{\alpha}_i)/2, \quad \text{if predicting mean},$$

$$\hat{a}_i = \sqrt{(\hat{\sigma}_i^2 + \hat{\alpha}_i' \hat{\Sigma}_v \hat{\alpha}_i)} \Phi^{-1}(p), \quad \text{if predicting } p\text{-th quantile},$$

for

$$\hat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^{T} (\log[(Q_{i,t}^k + 1)/V_{i,t}^k] - \hat{\alpha}_i' \hat{F}_t - X_{i,t}' \hat{\beta})^2$$

$$\hat{\Sigma}^v = \frac{1}{T-d} \sum_{t=d}^{T} \hat{v}_t \hat{v}_t', \quad \hat{v}_t = \hat{F}_t - \hat{\Phi}_1 \hat{F}_{t-1} - ... - \hat{\Phi}_d \hat{F}_{t-d}.$$

For each i, it obeys under the previous conditions,

$$\hat{a}_i - a_i = O_p(1/\sqrt{N} + 1/\sqrt{T}).$$

We now turn to proving the main assertion regarding the relative error. Under the stated assumptions and definitions, we see that

$$\begin{aligned} \text{relative error}_{i,t}^k &= \text{relative error}_{i,t} \\ &= \frac{|\hat{Q}_{i,t}^0 - Q_{i,t}^0|}{Q_{i,t}^0} \\ &= |\exp(-\epsilon_{i,t}) \exp(\hat{\alpha}_i' \hat{F}_t + X_{i,t}' \hat{\beta} + \hat{a}_i - \alpha_i' F_t - X_{i,t}' \beta - a_i) - 1|. \end{aligned}$$

We can further bound

$$\text{relative error}_{i,t} \leq \text{irreducible error}_{i,t} + \text{estimation error}_{i,t},$$

$$\text{irreducible error}_{i,t} := \frac{|\hat{Q}_{i,t}^{0,oracle} - Q_{i,t}^0|}{Q_{i,t}^0} = |\exp(-\epsilon_{i,t}) - 1|,$$

$$\text{estimation error}_{i,t} := \frac{|\hat{Q}_{i,t}^0 - Q_{i,t}^{0,oracle}|}{Q_{i,t}^0}$$

$$= \Big| \exp(-\epsilon_{i,t}) \Big(\exp(\hat{\alpha}_i' \hat{F}_t + X_{i,t}' \hat{\beta} + \hat{a}_i - \alpha_i' F_t - X_{i,t}' \beta - a_i) - 1 \Big) \Big|.$$

The last term can be upper-bounded, using the bounds on $\hat{\alpha}_i'\hat{F}_t - X_{i,t}'\hat{\beta} - \alpha_i'F_t - X_{i,t}'\beta$ stated in the main text:

$$|\hat{\alpha}_i'\hat{F}_t - X_{i,t}'\hat{\beta} - \alpha_i'F_t - X_{i,t}'\beta| = O_p(1/\sqrt{N}) + O_p(1/\sqrt{T}),$$

and the bounds on $\hat{a}_i - a_i$ stated above, as follows:

$$\exp(-\epsilon_{i,t})|\exp[O_p(1/\sqrt{T}) + O_p(1/\sqrt{T})] - 1| \leq \exp(-\epsilon_{i,t})|O_p(1/\sqrt{T}) + O_p(1/\sqrt{N})|.$$

This implies that for any finite constant K we have

$$\mathrm{E}[\text{estimation error}_{i,t} \wedge K] \leq K'\mathrm{E}[\exp(-\epsilon_{i,t})(1/\sqrt{T} + 1/\sqrt{N})],$$

where K' is another constant. Hence by the Markov inequality

$$\mathrm{P}(\text{estimation error}_{i,t} > t) \leq C_i(1/\sqrt{T} + 1/\sqrt{N}),$$

where C_i depends on t and K. Hence by the union bound

$$\begin{aligned}\mathrm{P}(\text{relative error}_{i,t} > c) &\leq \mathrm{P}(\text{irreducible error}_{i,t} > c/2) + \mathrm{P}(\text{estimation error}_{i,t} > c/2) \\ &\leq \mathrm{P}(\text{irreducible error}_{i,t} > c/2) + C_i(1/\sqrt{T} + 1/\sqrt{N}),\end{aligned}$$

and, substituting the expression $\text{irreducible error}_{i,t} = |\exp(-\epsilon_{i,t}) - 1|$, we obtain:

$$\mathrm{P}(\text{relative error}_{i,t} > c) \leq \mathrm{P}(|\exp(-\epsilon_{i,t}) - 1| > c/2) + C_i(1/\sqrt{T} + 1/\sqrt{N}).$$

Similarly we have that

$$\mathrm{E}[\text{relative error}_{i,t} \wedge K] \leq \mathrm{E}[\text{irreducible error}_{i,t}] + \mathrm{E}[\text{estimation error}_{i,t} \wedge K],$$

and the claimed bound follows from the above substitution.

Appendix B. A Robustness Check: Empirical Results for Agnostic Model for Active and In-Active Products. The Case of Electronics

Here we cut the panel data into two halves across the time dimension. We look at the first half of the panel, and isolate two subgroups of products (that existed in the first half of the panel):

- active products: products whose average unit sales were above 1;

- in-active products: products whose average unit sales were below .5.

Then we look in the second half of the panel and analyze the predictive equations for the probability of the Big Error Event for the two groups of products. In this section as well as in the remainder of the paper, we focus on the agnostic specifications as they tend to be more flexible and easier to interpret.

B.0.1. *Active Products.* We begin with descriptive statistics. Figure 20 shows the cross-sectional averages of Relative Forecast Error, Big Error Event occurrence, Age, and Number of Products, for each week $t = 1, ..., H$. Figure 21 shows the time averages of Relative Forecast Error, Big Error Event occurrence, Age, and Number of Products, for each product $j = 1, ..., M$.

From descriptive statistics we notice that the Relative Forecast Errors are bigger for the active products than for all products in the previous section. This literally means that the active products are harder to forecast, while in-active products (as we shall see below) are easier to forecast. We next estimate the agnostic model from the previous section on the active products and present the results in Table 4 and Figures 22 and 23. Relative to the previous analysis we see the following:

RC.1 For active products, the T effect is somewhat stronger than before.
RC.2 For active products, the N effect is essentially flat, as before.

There is not much to add here, and essentially the same comments apply as before, apart from the age effect being stronger here.

TABLE 4. Estimation Results of Fixed Effects Models for the Agnostic Model for Active Products

	Dependent variable:		
	1(Relative Forect Error $> X$)		
	Without Trend	Time Trend	Time Effects
	(1)	(2)	(3)
Age.Region.(20,58]	−0.030	0.010	−0.003
	(0.045)	(0.046)	(0.046)
Age.Region.(58,99]	−0.082	0.004	−0.006
	(0.051)	(0.053)	(0.053)
Age.Region.(99,148]	−0.184***	−0.048	−0.058
	(0.053)	(0.055)	(0.055)
Age.Region.(148,199]	−0.249***	−0.061	−0.070
	(0.055)	(0.059)	(0.059)
Age.Region.(199,278]	−0.309***	−0.060	−0.066
	(0.057)	(0.063)	(0.063)
N.Region.(1e+03,3.72e+03]	0.013	−0.077*	
	(0.043)	(0.046)	(0.000)
N.Region.(3.72e+03,7.74e+03]	0.032	−0.025	0.055**
	(0.038)	(0.039)	(0.022)
N.Region.(7.74e+03,9.65e+03]	0.035	−0.022	0.055
	(0.028)	(0.029)	(0.036)
N.Region.(9.65e+03,1.07e+04]	−0.001	−0.030	0.044
	(0.025)	(0.026)	(0.040)
N.Region.(1.07e+04,1.57e+04]			0.081*
	(0.000)	(0.000)	(0.047)
Trend		−0.565	
		(0.508)	
Squared.Trend		0.132	
		(0.309)	
Observations	88,149	88,149	88,149
R^2	0.237	0.239	0.242
Adjusted R^2	0.229	0.231	0.232

Note: Standard errors are clustered by product and date.

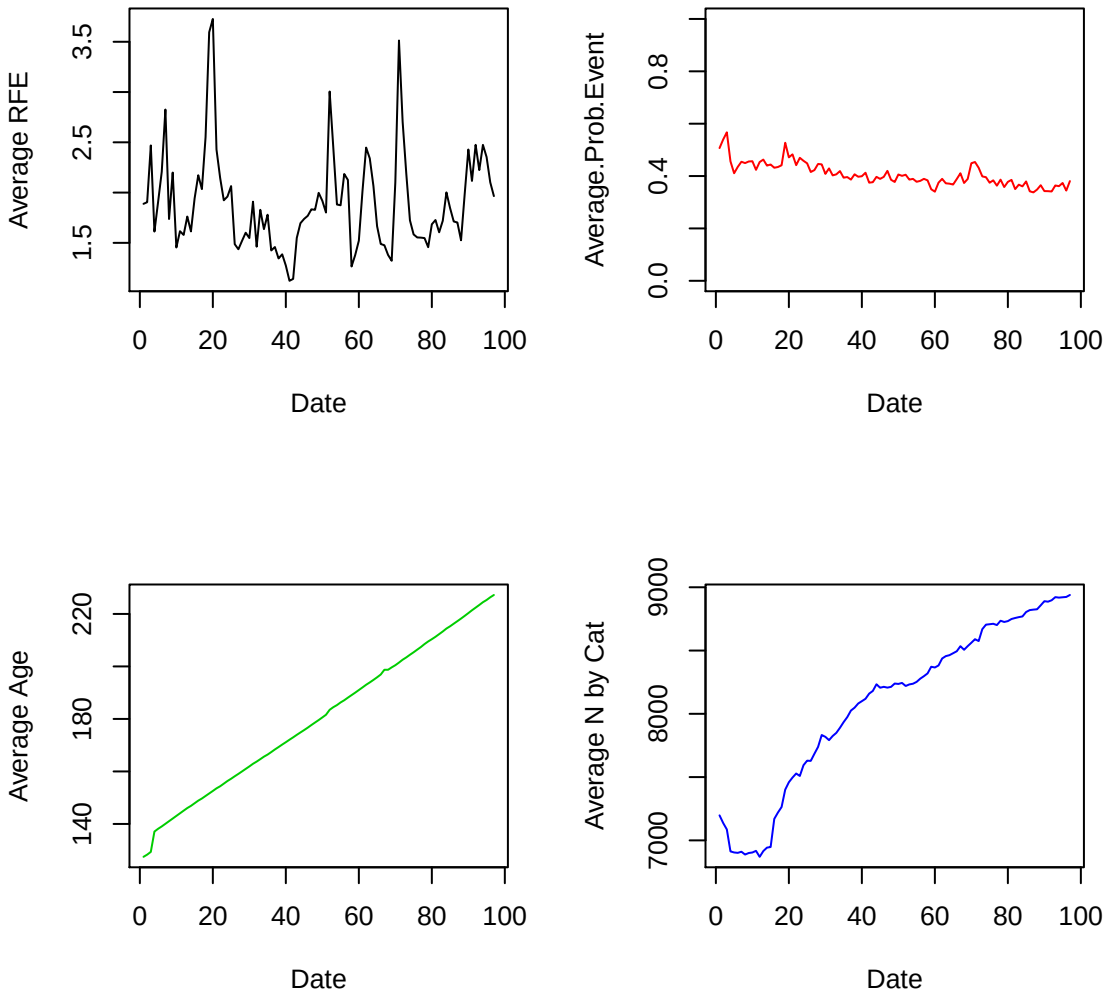


FIGURE 20. Active Products: Cross-sectional averages of Relative Forecast Error, Big Error Event occurrence , Age, and Number of Products, by Date

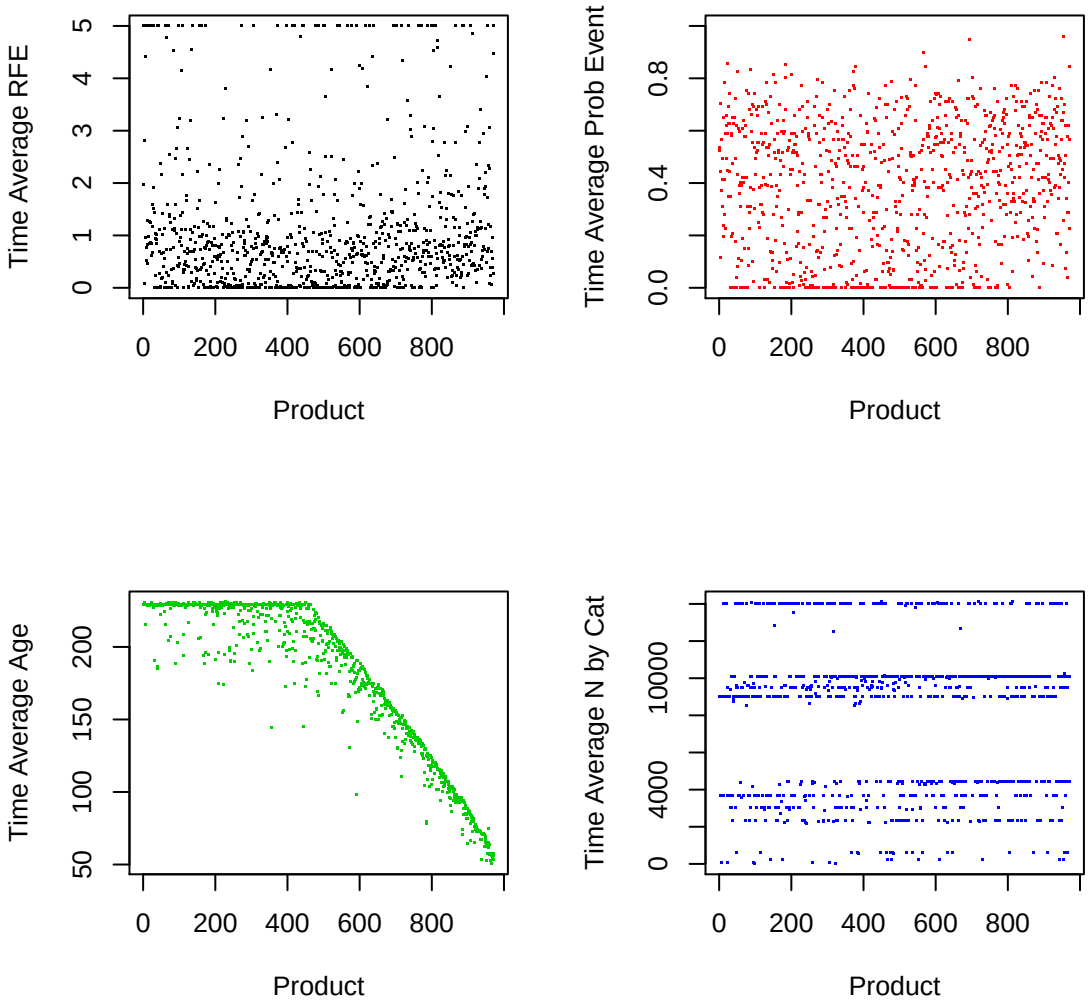
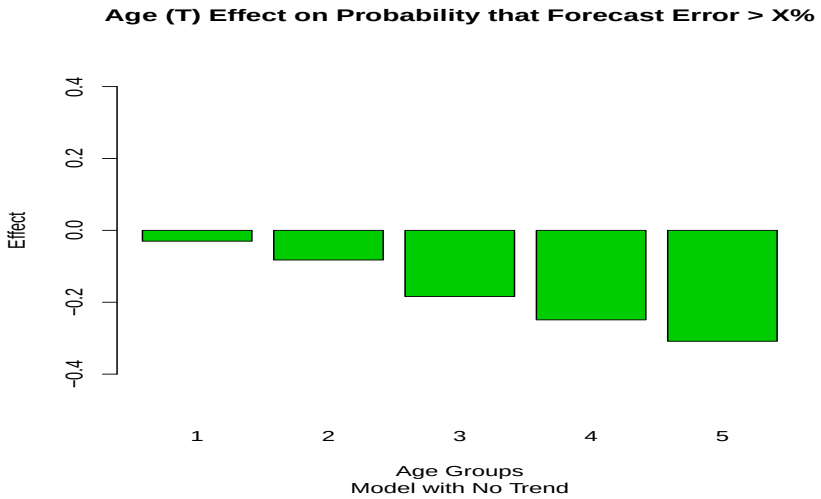
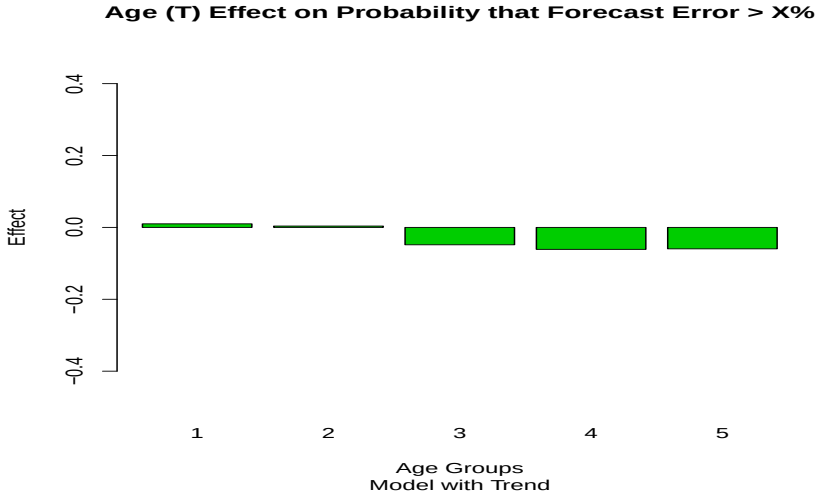


FIGURE 21. Active Products: Time Series averages of Relative Forecast Error (truncated at 5), Big Error Event occurrence , Age, and Number of Products, by Product





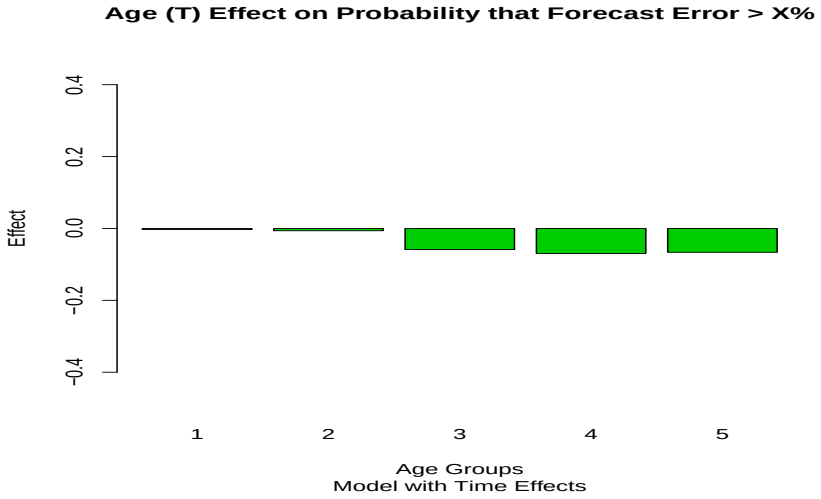
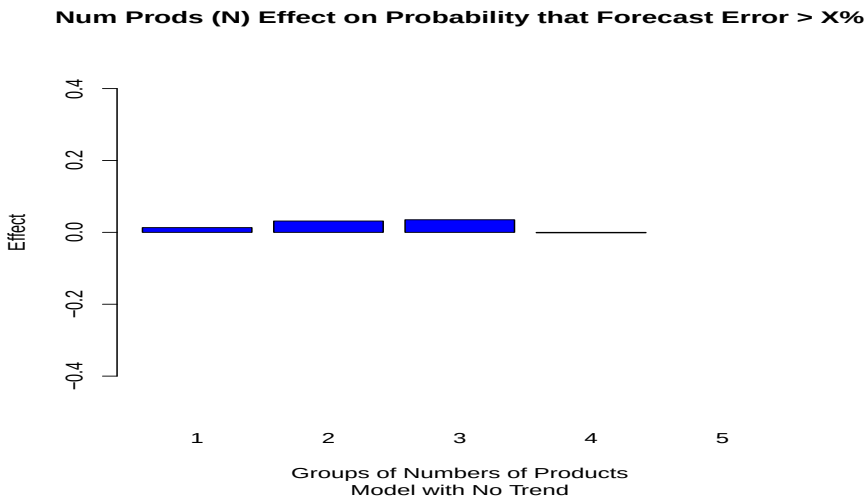
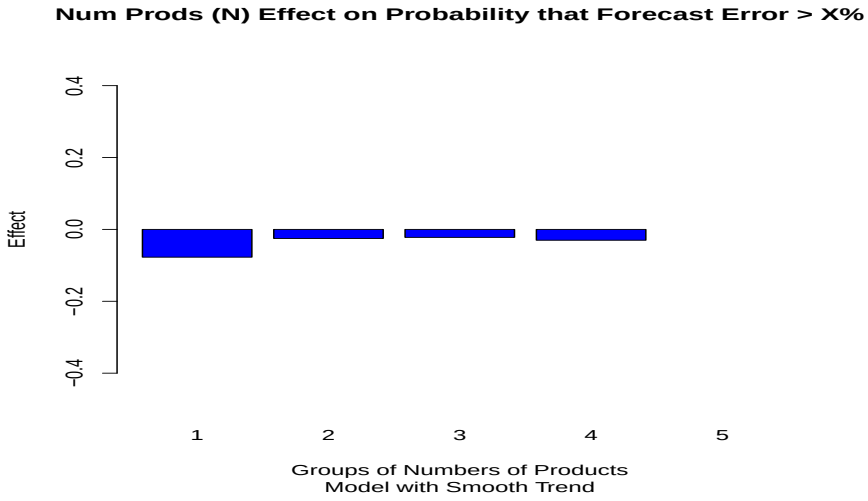


FIGURE 22. Active Products: The Estimated Impact of T on the Quality of Learning in the Agnostic Model with No Trend, Smooth Trend, and Time Effects





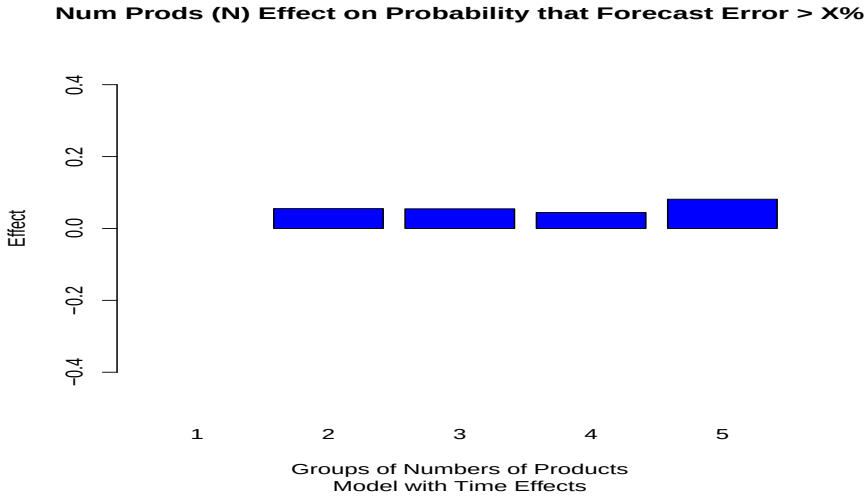


FIGURE 23. Active Products: The Estimated Impact of N on the Quality of Learning in the Agnostic Model with No Trend, Smooth Trend, and Time Effects

B.0.2. *In-Active Products.* We begin with descriptive statistics. Figure 24 shows the cross-sectional averages of Relative Forecast Error, Big Error Event occurrence, Age, and Number of Products, for each week $t = 1, ..., H$. Figure 25 shows the time averages of Relative Forecast Error, Big Error Event occurrence, Age, and Number of Products, for each product $j = 1, ..., M$.

From descriptive statistics we notice that the Relative Forecast Errors are much smaller for the active products than for all products in general. This literally means that the in-active products are easier to forecast, which agrees with any intuition as well as observations we made in Section 2. We next estimate the agnostic model from the previous section on the in-active products and present the results in Table 5 and Figures 22 and 23. Relative to the previous analysis we see the following:

RC.3 For in-active products, the T effect is somewhat less strong than before.

RC.4 For in-active products, the N effect is essentially flat, as before.

The result RC.4 agrees with Assertion 3 from comparative statics analysis in Section 2, namely that for in-active products the probability of Big Error Event is affected less by the size of data, which is the product age here.

Overall Conclusion from the Robustness Checks. Relative to the main analysis, we observe similar predictive impacts of data size on the probability of the Big Error Event, with the T effect being somewhat stronger for active products and somewhat weaker for in-active products. The effect of N continues to be essentially flat.

TABLE 5. Estimation Results of Fixed Effects Models for the Agnostic Model for In-Active Products

	Dependent variable:		
	1(Relative Forect Error $> X$)		
	Without Trend	Time Trend	Time Effects
	(1)	(2)	(3)
Age.Region.(20,58]	-0.069^{***}	-0.047^{*}	-0.027
	(0.026)	(0.027)	(0.028)
Age.Region.(58,99]	-0.110^{***}	-0.064^{**}	-0.043
	(0.027)	(0.027)	(0.029)
Age.Region.(99,148]	-0.133^{***}	-0.066^{**}	-0.043
	(0.028)	(0.028)	(0.030)
Age.Region.(148,199]	-0.153^{***}	-0.063^{**}	-0.037
	(0.029)	(0.029)	(0.031)
Age.Region.(199,278]	-0.184^{***}	-0.068^{**}	-0.041
	(0.030)	(0.031)	(0.032)
N.Region.(1e+03,3.72e+03]			-0.052^{**}
	(0.000)	(0.000)	(0.025)
N.Region.(3.72e+03,7.74e+03]	0.014	0.027	-0.027^{*}
	(0.018)	(0.018)	(0.015)
N.Region.(7.74e+03,9.65e+03]	0.020	0.043^{**}	-0.022^{**}
	(0.020)	(0.021)	(0.011)
N.Region.(9.65e+03,1.07e+04]	0.011	0.044^{**}	-0.021^{**}
	(0.021)	(0.022)	(0.008)
N.Region.(1.07e+04,1.57e+04]	0.022	0.055^{**}	
	(0.023)	(0.025)	(0.000)
Trend		-0.732^{**}	
		(0.290)	
Squared.Trend		0.356^{**}	
		(0.173)	
Observations	173,384	173,384	173,384
R^2	0.332	0.332	0.335
Adjusted R^2	0.323	0.324	0.326

Note: Standard errors are clustered by product and date.

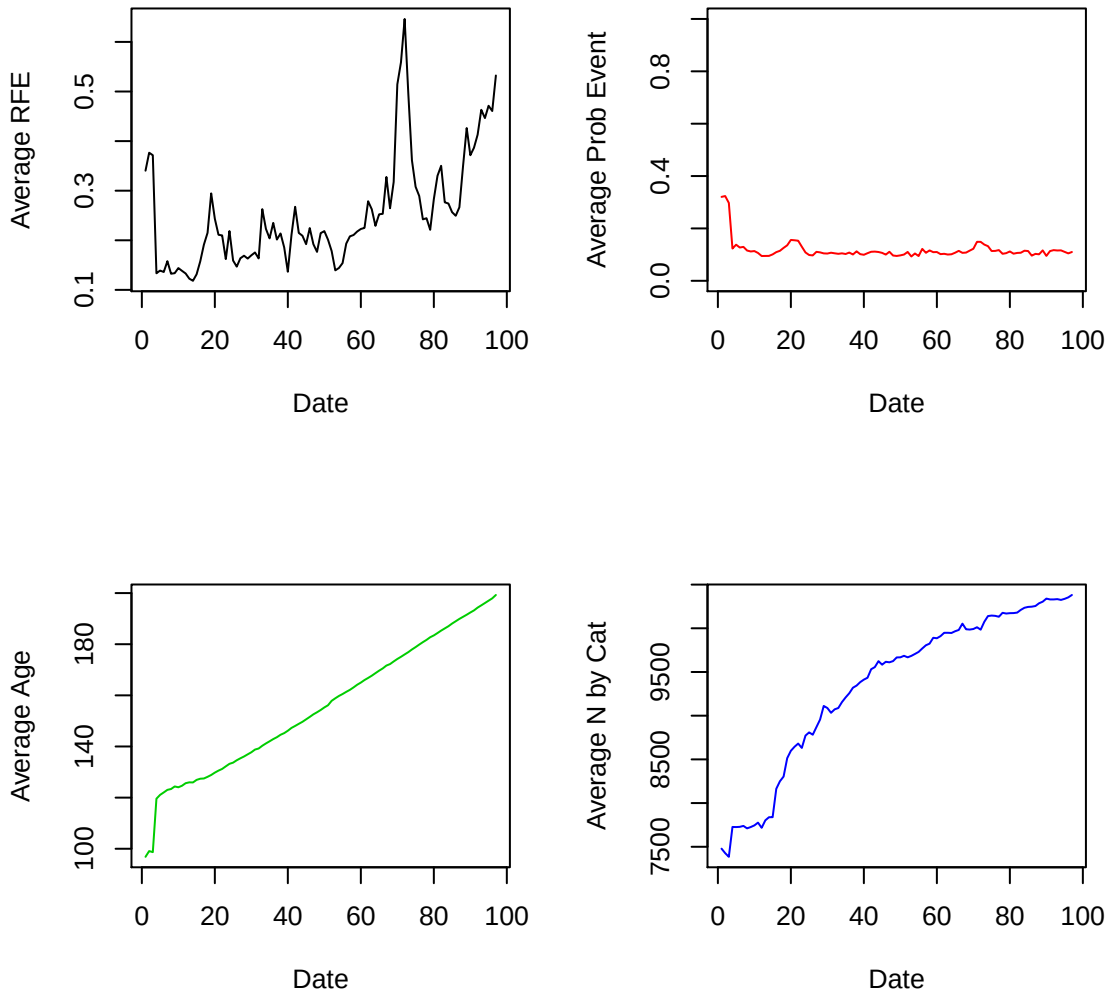


FIGURE 24. In-Active Products: Cross-sectional averages of Relative Forecast Error, Big Error Event occurrence, Age, and Number of Products, by Date

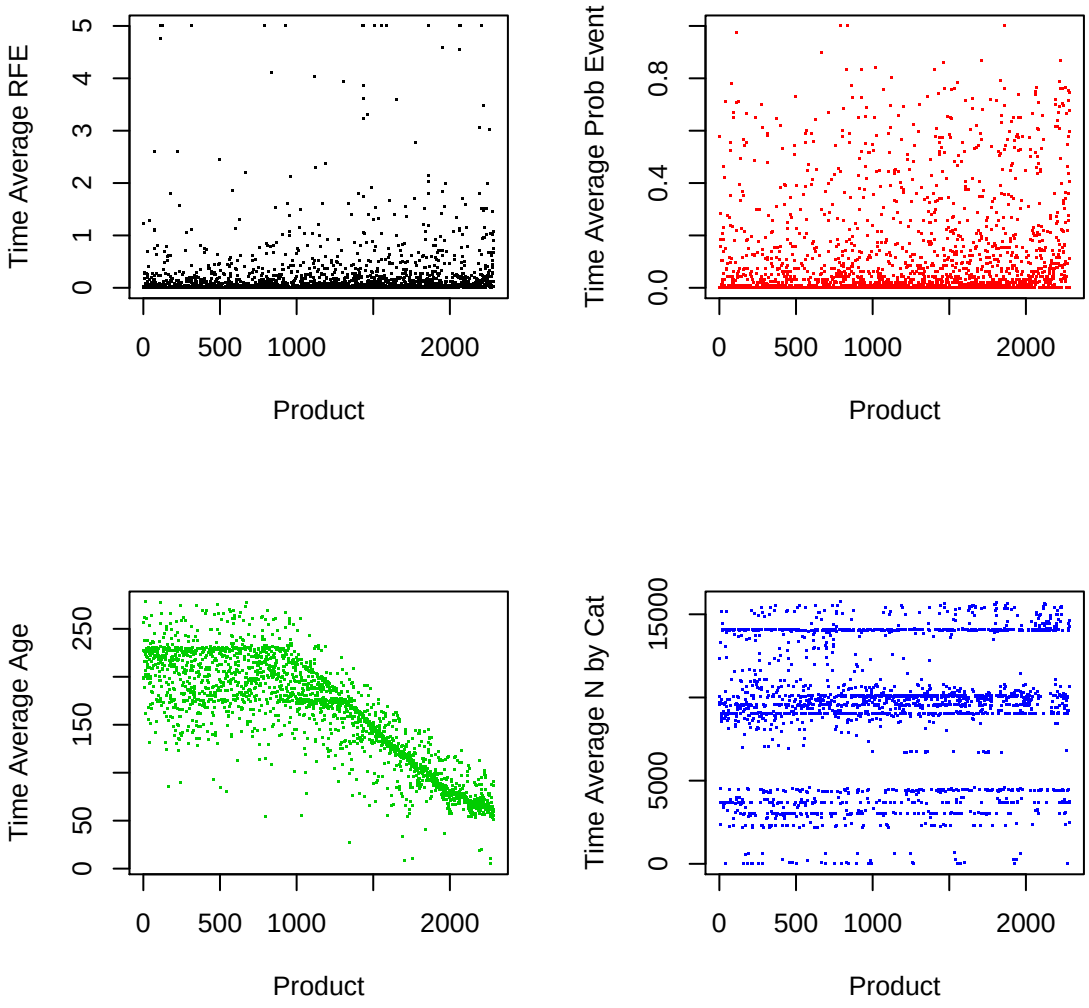
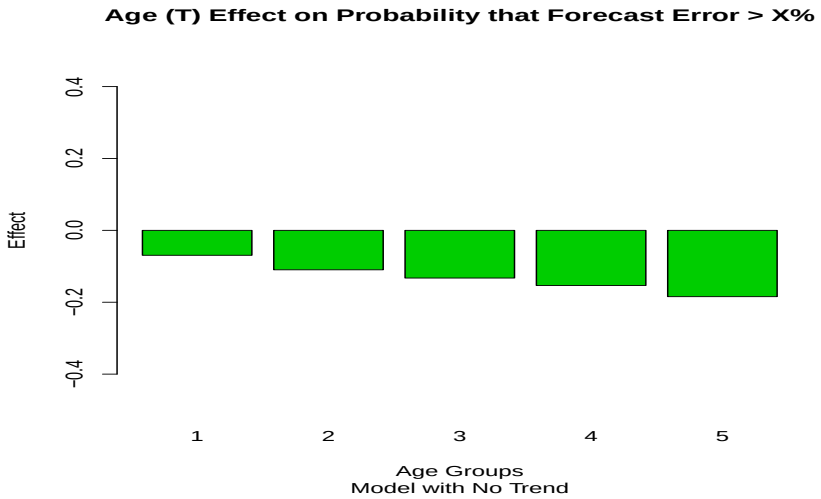
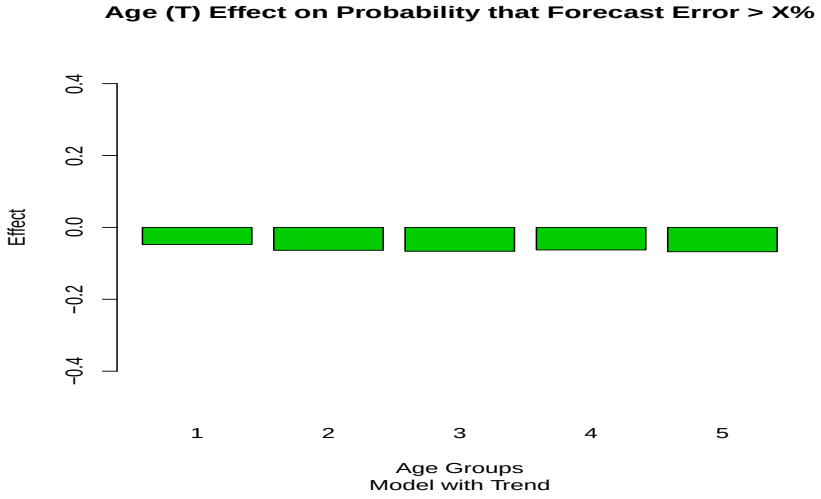


FIGURE 25. In-Active Products: Time Series averages of Relative Forecast Error (truncated at 5), Big Error Event occurrence, Age, and Number of Products, by Product





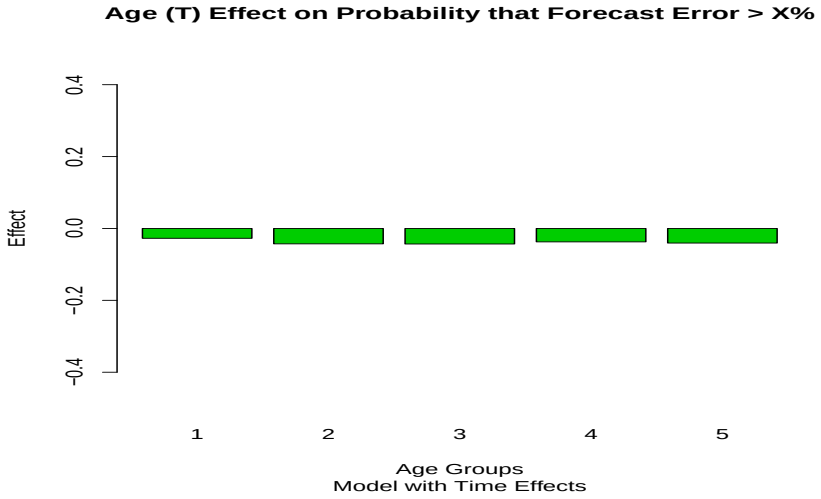
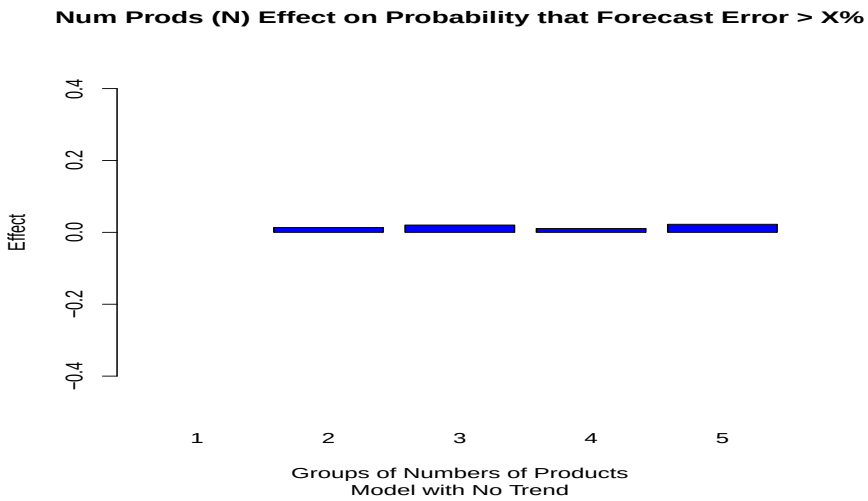
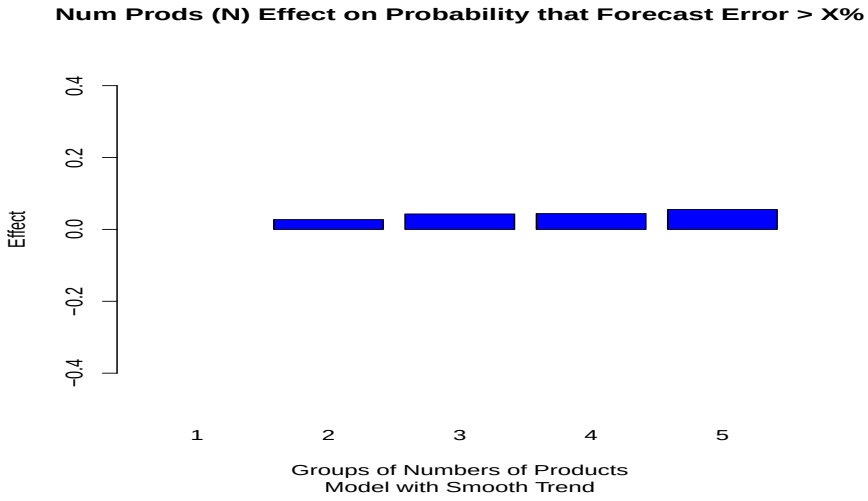


FIGURE 26. In-Active Products: The Estimated Impact of T on the Quality of Learning in the Agnostic Model with No Trend, Smooth Trend, and Time Effects





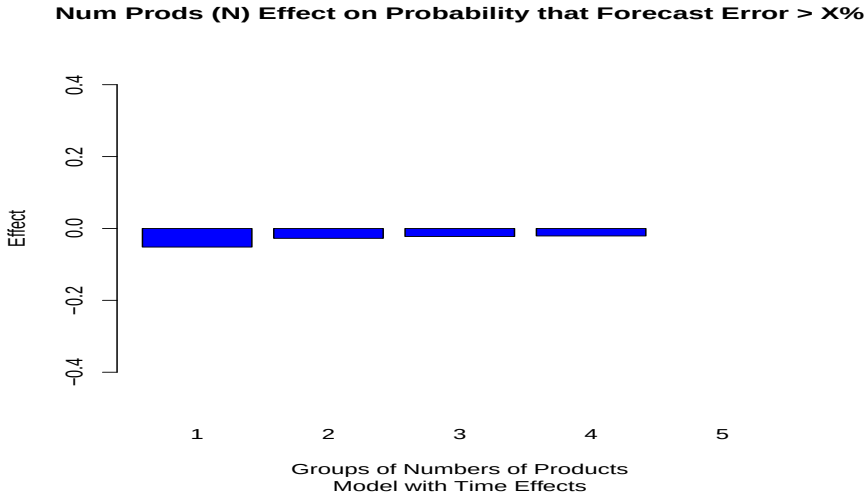


FIGURE 27. In-Active Products: The Estimated Impact of N on the Quality of Learning in the Agnostic Model with No Trend, Smooth Trend, and Time Effects

University of Washington

E-mail address: Bajari@uw.edu

MIT

E-mail address: vchern@mit.edu

University of Chicago

E-mail address: hortacsu@uchicago.edu

Amazon, Inc.

E-mail address: sjunich@amazon.com