A Multistate Model Approach to Producer Price Index (PPI) Response and Nonresponse Patterns

Joseph Valentine, U.S. Bureau of Labor Statistics, 4600 Silver Hill Rd, Suitland MD 20746

Monica Moleres, U.S. Bureau of Labor Statistics, 4600 Silver Hill Rd, Suitland MD 20746

Abstract

This paper analyzes survey nonresponse via the use of multistate Markov models. Typically, such models are used to describe a process or the evolution of a system over time. Using the response history of survey units (SU) initiated into the U.S. Producer Price Index (PPI) over a 10-year period, establishment respondents are shown to move in and out of various, mutually exclusive "states" of response and nonresponse. Using this paradata, a transition probability matrix is constructed and is then used to estimate transition probabilities at various points in a finite time horizon consistent with a seven-year sample. A subset of the paradata are then fed into a Random Forest model with specialized hyperparameters to predict the return of nonrespondents. The probability estimates from the Random Forest can then be used to determine which nonrespondents should be contacted with the goal of mitigating nonresponse overall.

Keywords: Nonresponse, Price Index, Markov Model, Machine Learning, Random Forest

This paper provides a summary of research results. The information is being released for statistical purposes, to inform interested parties, and to encourage discussion of work in progress. The paper does not represent an official BLS statistical data product or production series.

1. Introduction

The Producer Price Index (PPI) is a family of economic indicators that use voluntarily supplied survey data to measure the average change over time in selling prices received by domestic producers of goods and services. Because the PPI is a voluntary survey, establishment nonresponse is common when requesting a price for products, or items, at the individual level. An item-level analysis of these patterns can be performed via simple Survival Analysis with two states: response and nonresponse. However, not all items are the same, and because most respondents provide prices for more than one item, there are more factors to consider when attempting to understand respondent behavior at the aggregate level.

The purpose of this paper is to provide a comprehensive mapping of the states that PPI respondents enter and exit during their time in the PPI establishment survey, along with statistical estimates of each transition probability. The paper is organized as follows: part 2 provides background for the PPI and discusses the paradata used. A description of each state is followed by an example response history. Through construction of the response history, the calculation of time in each state is explained. Because respondents typically participate in the program for several years, this results in the construction of an unbalanced panel data set. Part 3 discusses methodology and estimation techniques used. Part 4 discusses the results and provides interpretation and analysis, while Part 5 discusses the results of a Random Forest model trained to estimate movements out of nonresponse status; Part 6 concludes with suggestions for mitigating nonresponse in the PPI program overall.

2. Background & Data

By law, most firms operating in the United States are required to participate in the Unemployment Insurance (UI) System. Due to its breadth and depth, the UI system serves as a basis for selecting potential survey respondents for the PPI. Typically, potential respondents are identified by their Employer Identification Number (EIN). After selection, a Bureau of Labor Statistics (BLS) representative attempts to solicit participation through in-person visit, a phone or video call. If the establishment agrees to participate, it provides pricing data for goods and services over time, the first instance of which occurs during the initiation meeting. For the purposes of this paper, specific goods or services are referred to as items and the number of items for which an establishment provides data is called their burden. After the initial visit, respondents are contacted to provide data on a previously agreed-upon schedule. This is referred to as "repricing". Typically, respondents provide data monthly, though some follow an alternative schedule, such as quarterly. Throughout the repricing process, establishments enter and exit different, mutually exclusive repricing states defined in Table 1 below.

As explained in Table 1, for a given month, respondents are in one of six repricing states based on the number of items they have or have not provided data for. Since the paradata used in the analysis below is taken from outside of the four-month correction period for BLS data, it is to be considered fixed or immutable. In the first state, the respondent has provided data for all items requested—full response. In the second state, where the respondent provides a partial response, the number of items for which prices are provided is less than the full burden, but no items are considered in nonresponse status. An item has entered nonresponse status only when the respondent does not provide prices to PPI when requested. This nonresponse prompts a follow-up from an analyst via phone. If prices are received over the phone, it is considered a response. If the analyst cannot connect with the respondent, another call is made. This procedure is typically followed for up to 12 consecutive unanswered requests, via email, at which point the given item is discontinued or deleted. The third state, when a respondent is off-cycle occurs when a respondent does not provide prices because no response was scheduled. The fourth state, partial nonresponse, is entered when the number of items in nonresponse status are less than the total respondent burden and no items are in response status. State five, full nonresponse, occurs when the number of items in nonresponse status exactly matches the respondent burden. Finally, deletion, at the establishment level, occurs when all items from a given establishment meet the threshold for deletion. These concepts are illustrated below in Table 2 and Table 3, which provide generic examples of how establishments can move through the repricing states discussed above.

Table 1: PPI Repricing States and Definitions		
State	Name	Definition
1	Full Response (FR)	Items with data provided equals respondent burden
2	Partial Response (PR)	Items with data provided are non-zero, but less than burden and no items are in nonresponse status
3	Off Cycle (OC)	No data requested; nonresponse is not possible unless data is not provided in most recent months when request was made; data may still be provided; if data are provided it is considered a response
4	Partial Nonresponse (PNR)	Number of items in nonresponse status less than burden but not zero; no items in response status
5	Full Nonresponse (FNR)	Number of items in nonresponse status equals respondent burden
6	Deletion (D)	Number of deleted items equals respondent burden
Note: Respondent burden is the number of items a given establishment provides data for. An item is in nonresponse status if data is not provided when requested. Once items enter nonresponse status, they remain until data are provided.		

Table 2: Example of Item-Level Repricing History									
SU	item	irm	rpc.length	cycle	price	response	cum.response	nr	cum.nr
D056	A01	202109	1	M	22.9	1	1	0	0
D056	A02	202109	1	M	75	1	1	0	0
D056	A01	202110	2	O	22.9	1	2	0	0
D056	A02	202110	2	O		0	0	0	0
D056	A01	202111	3	O		0	0	0	0
D056	A02	202111	3	O		0	0	0	0
D056	A01	202112	4	M	13.6	1	1	0	0
D056	A02	202112	4	M	90	1	1	0	0
D056	A01	202201	5	O		0	0	0	0
D056	A02	202201	5	M	19	1	1	0	0
D056	A01	202202	6	O		0	0	0	0
D056	A02	202202	6	M		0	0	1	1
D056	A01	202203	7	M		0	0	1	1
D056	A02	202203	7	M		0	0	1	2
D056	A01	202204	8	O	17	1	1	0	0
D056	A02	202204	8	M	91	1	1	0	0
D056	A01	202205	9	O	18.75	1	2	0	0
D056	A02	202205	9	M	93	1	2	0	0
Note: irm is index reference month. response & nr (nonresponse) are indicator variables equal to one if a price is and isn't reported, respectively, in the given index reference month; cum.response & cum.nr are cumulative number of months with prices reported and not reported, respectively, through the current irm. Cycle is the repricing cycle of the item. M indicates that a price was requested. Rpc.length indicates the time, in months, the item has spent in the PPI survey.									

In Table 2 above, the survey unit (SU) and item columns denote the business establishment and the particular item being repriced, respectively. The irm variable represents the month for which a given price would be provided, rpc.length provides a numeric representation of how long the respondent has been providing prices for the particular item. The cycle variable is an indicator of whether or not the respondent was expected to provide data. The variable takes one of several values (e.g. "O", "M", "N"), each of which has a specific meaning with respect to response and nonresponse status. A cycle value of "M" means that the PPI requested data in a given month, "O" indicates the item in particular is off-cycle, and "N" indicates that no data was requested for the given month. In this specific case, both items are on irregular, or atypical repricing schedules; repricing is more often conducted periodically on a monthly, quarterly, or annual basis. The price variable represents the price provided by the respondent. In this case, since the data are synthetic it is purely hypothetical, although each observation does reflect common trends in repricing. The response and nr variables indicate whether or not data was provided in a given month, while cum.response and cum.nr are cumulative totals of consecutive months in response and nonresponse status respectively.

In the example above, establishment D056 provides prices for two items, A01 and A02, both on an atypical basis. For the first month in question, September 2021, data is provided for both items as requested. In October, however, data are only provided for A01. Although no data is provided for A02, it is not in nonresponse status because the item is off-cycle and data were not requested. October (202110) provides an important distinction regarding response and nonresponse states: because data was provided for item A01, it is considered a response, which means that since item A02 is off-cycle and no data was provided the SU is considered to be in *partial* response (state 2). In this particular month, the number of responses (one) is less than the total number of items being repriced (two). And since no items are considered in nonresponse status, it is coded as a partial response. Both partial response (state 2) and partial nonresponse (state 4) states exist to account for special "edge" cases that occasionally occur within the repricing history of a given SU.

Both items are off-cycle for November as well. No data is requested again for either item until December, when data are requested and provided for both items, but only data for A02 are requested and provided for the following month, January 2022. No data is provided for either item in February 2022. In this case, the SU is in partial nonresponse rather than full nonresponse. This is because only item A02 is in nonresponse status due to data being requested in the given month; item A01 is off-cycle and so no data was requested, therefore the item A01 cannot be in nonresponse status, placing the SU in state 4—partial nonresponse. In March of 2022 data are requested for both items, but no data is provided. This is the first time either item enters nonresponse status. Consequently, it is the only month in the time span considered in which the SU is in full nonresponse status. In both April and May, data are provided for both items so the SU is in full response for both months.

To understand how the data in Table 2 relate to Table 1, it is necessary to aggregate, or "roll up" the data in Table 2 to summarize it so that it represents what repricing state the establishment is in and not simply the items representing the establishment. This is demonstrated in Table 3 below, which summarizes Table 2 but at the aggregate level.

Table 3 takes a top-down view of the information in Table 2, demonstrating how the establishment in question moves through various states discussed in Table 1. Each line in Table 3 corresponds to two lines in Table 2, since in this example there are two items being repriced. When price data are requested, the set of possible states include full.resp, part.nr or full.nr. When price data are off-cycle, the set of possible

states includes full resp, part.resp, or off.cycle, which is summarized neatly in the state column in Table 3. For the first month in question, they are in full response status, or state 1, as shown by the first two lines in Table 2. In the second month, they exit state 1 and enter state 2, since data are not provided for A02. Since A02 is considered off-cycle in October 2021, it is not in nonresponse status so establishment D056 is in state 2—partial response. In the third month, both items are off-cycle and no data are provided so the establishment is in state 3, off-cycle. The fourth month, December 2021, sees the respondent return to state 1, which they exit in month 5 to enter state 2, and enter state 4 in month 6. In month seven, March 2022, they then enter state 5, full nonresponse, since data are requested but not provided for either item. They then move to state 1 for the remainder of the period in question.

Table 3: SU Level Repricing History

SU	burden	irm	rpc.length	full.resp	part.resp	off.cycle	part.nr	full.nr	state
D056	2	202109	1	1	0	0	0	0	1
D056	2	202110	2	0	1	0	0	0	2
D056	2	202111	3	0	0	1	0	0	3
D056	2	202112	4	1	0	0	0	0	1
D056	2	202201	5	0	1	0	0	0	2
D056	2	202202	6	0	0	0	1	0	4
D056	2	202203	7	0	0	0	0	1	5
D056	2	202204	8	1	0	0	0	0	1
D056	2	202205	9	1	0	0	0	0	1

Table 3 is informative because it shows how easily respondents move in and out of each state. Overall, what Table 3 shows is the complexity of the repricing process for a given respondent. More specifically, it shows that there are both random and deterministic elements to the movements of PPI survey respondents after initiation. Because each respondent enters the survey in the same place, state 1, and because there are very few strict rules governing their movements, the repricing process is essentially a stochastic one—which makes it an ideal candidate for applying Markov or semi-Markov models.

3. Methodology

A Markov chain describes a sequence of possible events in the evolution of a stochastic process. Though capable of producing several numerical outputs, the primary goal is to estimate the probability of said events, as described below.

Let $t \geq 0$ be a parameter representing time in a set T, and let $X(t)$ be a random variable for every $t \in T$. The collection of random variables $R = \{X(t), t \in T\}$ therefore represents a stochastic process, which is observed over time. The values taken by X at various points in time t represent the individual states that the process inhabits over time, denoted by $S = \{1, 2, 3, 4, 5, 6\}$ and summarized in Table 1 above. In this case, $X(t)$ can be said to represent a given survey unit at time t, measured in months, and its value is represented by its repricing state. More specifically, what this implies is that the PPI repricing process R is defined by a set of respondents and their behavior over time.

Because the state space S is discrete, the stochastic process R is considered a Markov chain such that the following holds:

$$Pr(X_{t+1} = x_{t+1} | X_t = x_t, \ldots X_0 = x_0) = Pr(X_{t+1} = x_{t+1} \mid X_t = x_t) \quad (1)$$

What equation (1) states is that given all available information in the present, the future is independent of the past. In other words, the state that X inhabits in period $t+1$ is affected only by the data available in period t. Equation (1) is important because it governs the movement of X between various states over time. When applied to paradata resembling Table 3, the property defined in equation (1) is satisfied because the information regarding what state X is in and the amount of time spent in each state is updated and summarized in the current period, so only data in period t is needed to understand the transitions X may make in period $t+1$.

These transitions are represented via the following conditional probability:

$$p_{jk} = P(X_{t+1} = k \mid X_t = j), j, k \in S \quad (2)$$

Equation (2) gives the probability of a given survey unit moving from state j to state k in one step, or one time period. The probability p_{jk} is an element, from row j and column k, in an n by m nonnegative matrix $P(t)$ called the transition probability matrix.

The matrix $P(t)$ is essential to understanding respondent behavior over time because it can be used to estimate several important quantitative measures that provide insight into the repricing process and its evolution. One of these measures is the total amount of time spent in each of the six repricing states. This is given by

$$L_s = \int_{t_1}^{t_2} P(t)_{r,s} dt \quad (3)$$

Where r is the state at the start of the repricing process, in this case 1, and s is the state in question. Equation (3) is estimated via numerical integration. What it amounts to is the expected amount of time spent in state s between initiation, t_1, and some specified future time t_2. Other useful estimates include the expected number of times a survey unit will enter each state, the expected first passage into each state, the average time spent in each state per entry.

4. Results

To estimate $P(t)$ initial values should be supplied. For the process R, the initial values are derived from actual transitions observed over the lifecycle of all PPI respondents initiated between January 2014 and December 2024. The transitions are discussed in further detail below.

$$M = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 482350 & 1985 & 3808 & 3097 & 5301 & 22 \\ 2 & 1373 & 6551 & 894 & 687 & 38 & 0 \\ 3 & 2204 & 822 & 14743 & 243 & 1220 & 1 \\ 4 & 3265 & 289 & 126 & 28728 & 1363 & 0 \\ 5 & 4362 & 8 & 280 & 1558 & 52145 & 176 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\rightarrow M^* = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0.971 & 0.004 & 0.008 & 0.006 & 0.011 & 0.000 \\ 2 & 0.144 & 0.686 & 0.094 & 0.072 & 0.004 & 0.000 \\ 3 & 0.115 & 0.043 & 0.767 & 0.013 & 0.063 & 0.000 \\ 4 & 0.097 & 0.009 & 0.004 & 0.851 & 0.040 & 0.000 \\ 5 & 0.075 & 0.000 & 0.005 & 0.027 & 0.891 & 0.005 \\ 6 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}$$

Consistent with the state space S, matrices M and M^* are 6×6 representations of the movements that have occurred in R between January 2014 and December 2024. Matrix M provides the raw number of transitions between states, while M^* produces relative proportions that may be interpreted as probabilities. In both M and M^*, rows represent exit states and columns represent entry states. For example, the entry in row 1 and column 1 of M represents the number of times an SU was in full response status and remained there the following month, while $M^*\ (1, 5)$ represents the probability of a respondent going from full response status to full nonresponse status. Note that each row of M^* sums to 1. Using M^*, $P(t)$ may be estimated at various values of t via Maximum Likelihood.

4.1 Aggregate Results—Full Response Status

Figure 1, Transition Probabilities From Full Response Status

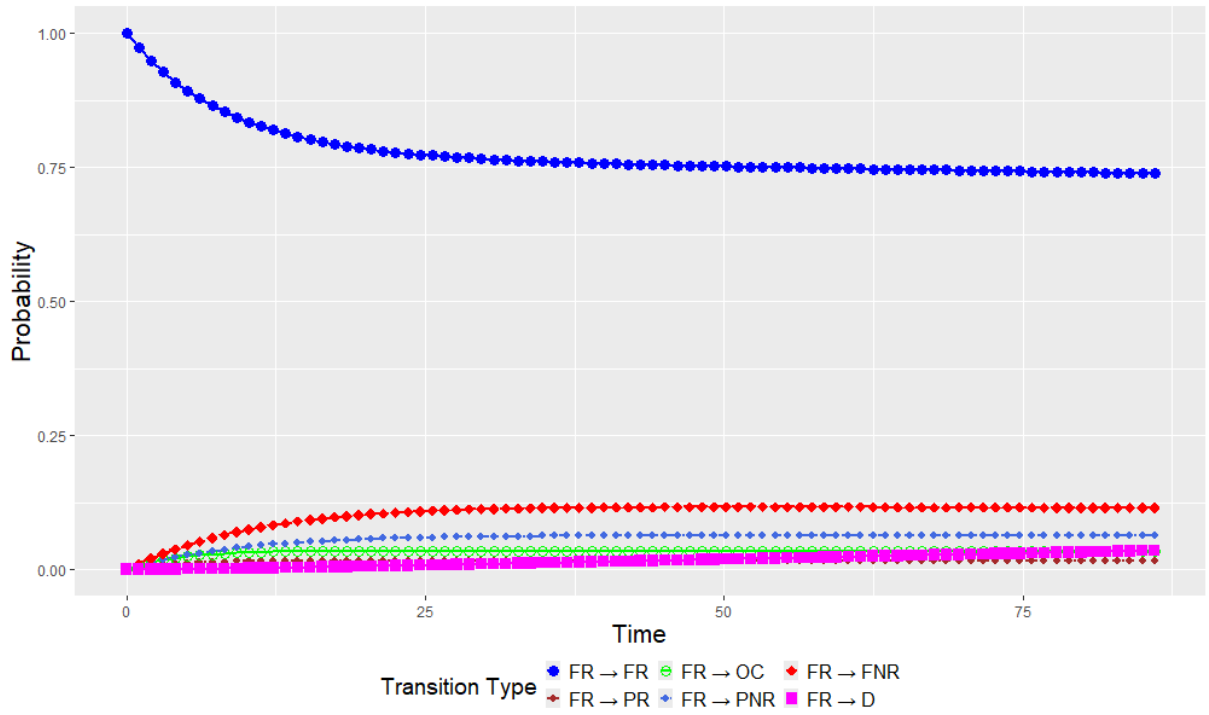


Figure 1 provides a graphical representation of the first row of $P(t)$ over 84 months, equivalent to a standard PPI Sample. What it shows is the probability that a respondent in full response status will be in another state at a given time t on the x-axis. The solid blue line with filled circles near the top of the x-y plane shows the probability that respondents will remain in full response status. It shows that the most productive respondents are much more likely to continue providing a full set of data for the PPI throughout the period under consideration. The overall profile suggests an initial drop-off, as some respondents exit at a steep rate, followed by a slowdown in the likelihood of exiting full response. The likelihood of a respondent

exiting full response to enter full nonresponse status is the second most likely outcome, indicated by the dashed red line with filled circles. It is important to remember that all respondents start in full response status due to the initiation visit, which is why all other states start at zero. Over time, the data suggest that in the early months after initiation, the likelihood of respondents entering full nonresponse status increases rapidly, until around their 24th month in the PPI system, after which the likelihood of their entry into state 5 essentially flattens. Overall, what Figure 1 suggests is that once respondents are initiated, the likelihood of nonresponse increases the most in the early months in the system. Another takeaway is that the estimated probability of a respondent remaining productive 84 months after initiation is around 74 percent. This does not imply that a respondent will never become nonproductive by moving into full or partial nonresponse status, but rather that, 84 months after initiation, they are likely to remain fully productive. Figure 1 also shows that, as the probability of moving into full nonresponse flattens, the probability of going from full response to deletion increases steadily, as indicated by the solid pink line with squares. Since this curve represents the likelihood of a respondent going from fully productive to no longer being in the PPI system, it can be thought of as a measure of how likely it is that a productive respondent directly "opts out" of participation. This is often accomplished through a direct request to a PPI analyst via phone or email. Toward the end of the period under consideration, deletion becomes the third most likely outcome for a productive respondent. This could be due to various reasons, including but not limited to, respondent fatigue, changing responsibilities at the SU level, or an SU going out of business.

4.2 Aggregate Results –Full Nonresponse Status

Figure 2, Transition Probabilities From Full Nonresponse Status

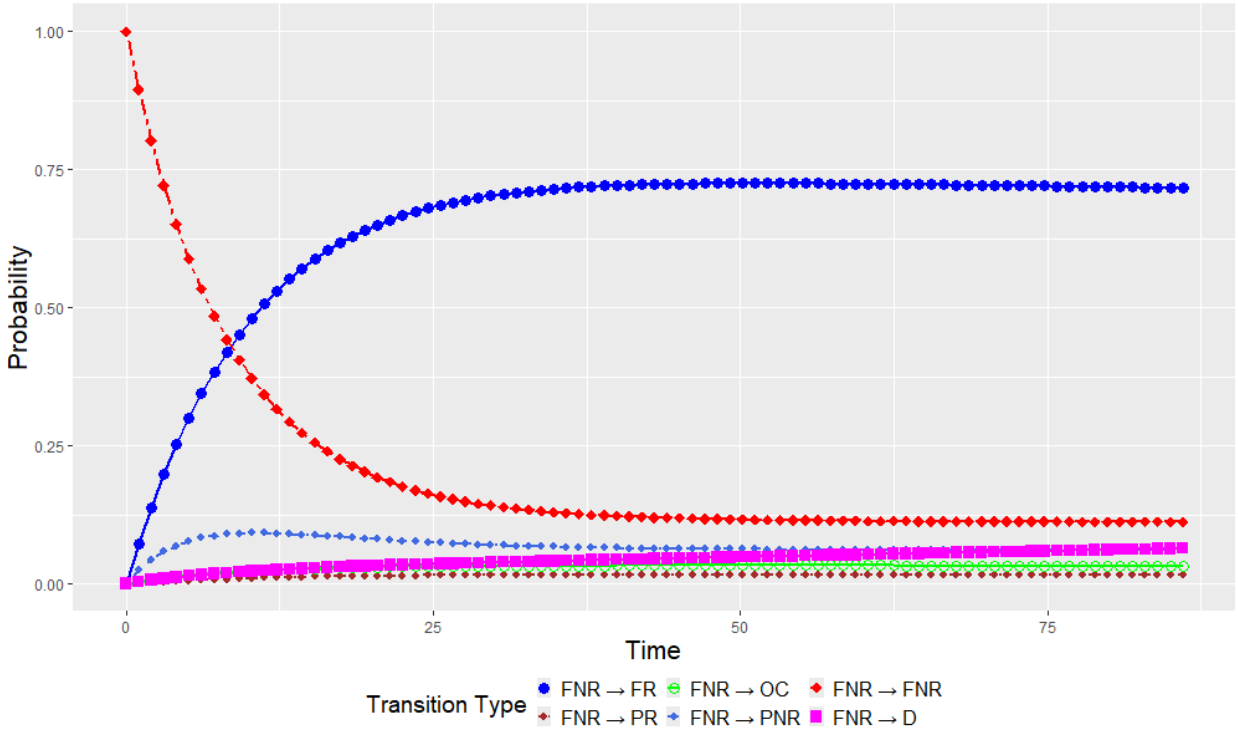
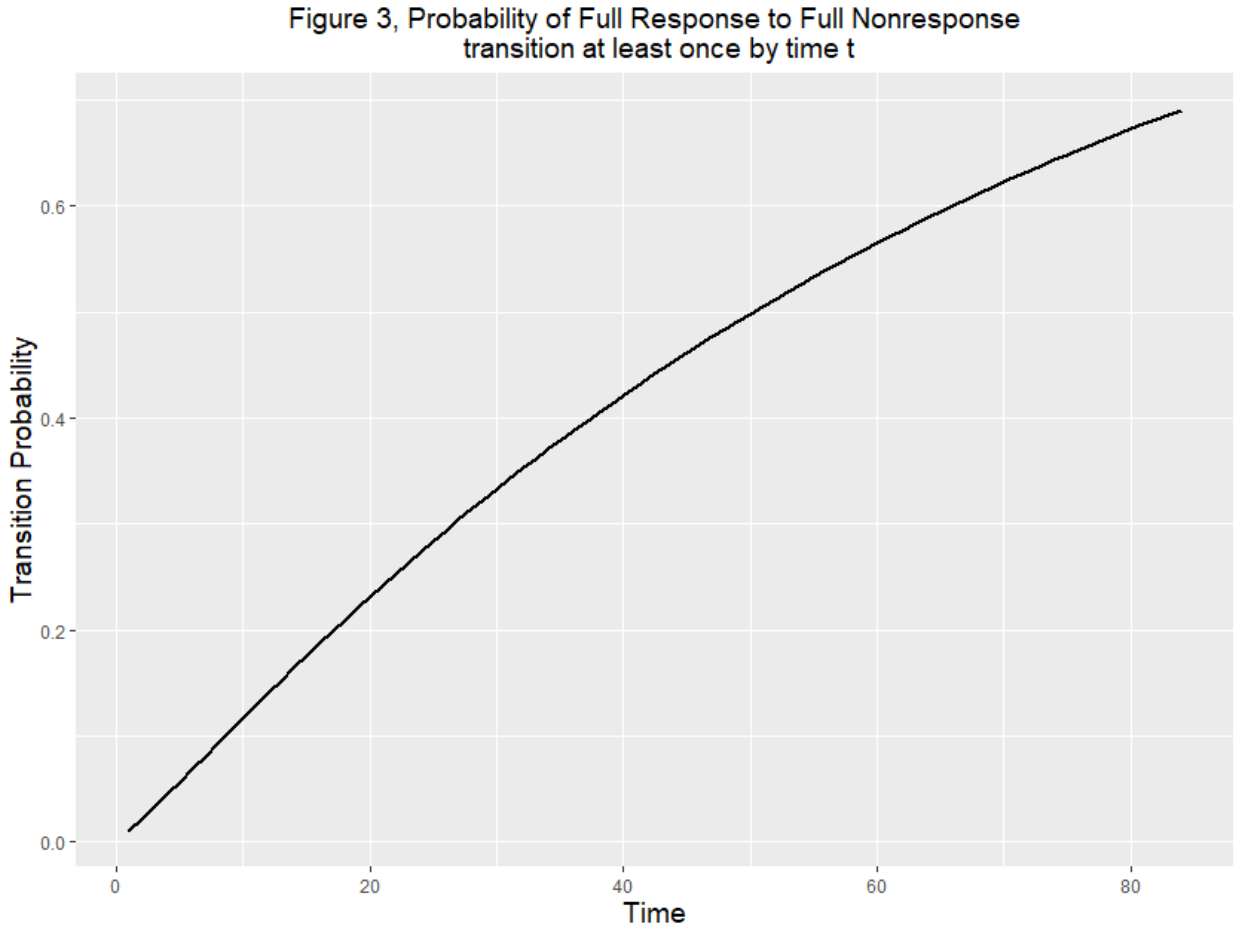


Figure 2 above may be considered the nonresponse analog to Figure 1. It provides graphical depiction of row 5 of $P(t)$ through 84 months. The starting point is the dashed red line with filled red circles, which gives the probability that an SU that is currently in full nonresponse status will be there t months in the future. The probability of an SU leaving full nonresponse status to enter full response status t months in

the future is given by the solid blue line with filled circles. Put another way, the solid blue line with filled circles gives the probability that establishments who are providing no data will be providing all the data requested, equal to their burden, t months from now. The dashed red curve gives the probability that nonrespondent establishments will continue to provide no data t months into the future. As the dashed red line shows, the likelihood that a nonrespondent will continue to provide no data declines sharply at first, following a negative logarithmic pattern, almost becoming flat around 25 months. Similarly, the solid blue curve with circles rises sharply at first, indicating that although a multitude of respondents do in fact enter nonresponse after initiation, they soon enough begin to provide data. This is not uncommon and is likely due to the lag between initiation and first repricing: after initiation, there is typically a period of several months until a new respondent is fully established in the monthly survey. During this time, emails often appear in spam folders and in some cases, respondents may not recall agreeing to provide data for PPI. Figure 2 also shows that over time, the likelihood of deletion increases somewhat linearly. Comparing the solid pink line with squares in Figure 2 to the one in Figure 1 reveals that the rate at which respondents are deleted is greater for those exiting full nonresponse status than it is from those leaving full response status. What is not shown is that, Asymptotically, the likelihood of deletion overtakes the likelihood of remaining in nonresponse status. This is because after a long enough period of sustained nonresponse the SU is likely to be deleted.

Overall, what should be taken away from Figures 1 and 2 is that by and large, the default position of the repricing process R is largely one of inertia: on a month-to-month basis, where an establishment will be next month is largely determined by where they are in a current month. Those in full response status tend to stay there, and those that are in full nonresponse tend to as well. This is observed in the intermediate states of partial response and nonresponse as well, although those cases represent a small share of the overall number of transitions that have occurred during the time period in question. They have been included to provide as robust a picture of the repricing process as possible.



In terms of understanding *when* entries into full nonresponse status occur, and how likely they are, Figure 3 provides an auxiliary representation of the movement from full response to full nonresponse over time. Figure 3 shows the probability that an SU will have transitioned from full response to full nonresponse *at least once by time t*, given on the X-axis. Note that this is different from the solid blue curve with circles in Figure 1which gives the probability of transition from state 1 to state 5 at specific time points in the future. The curve in Figure 3 gives the probability that the transition will have already occurred by time t. This does not mean that the SU will be in full nonresponse at time t but simply that it will have been there at least once by that time. For example, there is a 14 percent probability that a given respondent will have entered full nonresponse status at least once by their 12^{th} month in the PPI system. Figure 3, combined with Figures 1 and 2, provides a full representation of response and nonresponse in the PPI system.

Table 4: Total Time Spent in Repricing States		
State	Name	Time (Months)
1	Full Response	65.35
2	Partial Response	1.29
3	Off Cycle	2.77
4	Partial Nonresponse	4.73
5	Full Nonresponse	8.74
6	Deletion	1.22
Note: estimates from numerical integration, based on 84-month time period		

A useful measure, for the purposes of sample planning, involves how much data, and of what quality can be obtained for a given sample length. Table 4 above provides estimates as the result of equation (3) via Numerical Integration. As the table shows, for an 84-month time horizon, a given SU is expected to spend about 65 months total in full response status, roughly 9 in full nonresponse status, and about 1 month in deletion status. This amounts to a full response for almost 8 of every 10 months in the PPI system and about 1 in 10 months in full nonresponse status[1].

5. Extensions

The results above provide significant insight into the behavior of PPI respondents at the aggregate level. Providing structure to the data shown in Tables 2 and 3 via a multistate Markov model allows the repricing process R to be analyzed systematically at the aggregate level. Because the plots in Figures 1 and 2 are probabilistic in nature, the data behind them can be used to predict the behavior of respondents.

The behavior of nonrespondents is of particular interest. One major takeaway from the above analysis is that respondents typically provide all requested data or none at all. This creates a binary problem that requires maximizing transitions from full nonresponse to full response and minimizing transitions from full response to full nonresponse. One potential approach to achieving this outcome is to employ the use of Machine Learning (ML) tools.

The goal of using Machine Learning in this context is to predict which nonrespondents are more likely to move to full response in the following month. If it is possible to know which respondents are most likely to return, targeted efforts can be made to contact them to ensure data is provided the following month. To this end, a Random Forest classifier was employed. The classifier was trained on data sampled using a modified sampling technique. To prevent data leakage, all SUs initiated between 2014 and 2021 were used for training and testing, while SUs initiated between 2022 and 2024 were kept separate as an unlabeled holdout set. The model was trained using several variables, defined in Table 5 below.

[1] A useful follow-up question would entail a breakdown of how these estimates have changed over time for the PPI at large. This investigation is currently underway.

Table 5: Variable Definitions and Importance		
Variable	**Definition**	**Importance**
nr_exit_count	Number of exits from full nonresponse status	100
time	Total time in PPI, in months	48.89
size_measure2	Size measure 2 (typically quarterly wages)	26.62
coll_shipments_receipts	Collected shipments and receipts	19.4
su_weight	Sum of item-level index weights	17.9
size_measure1	Size measure 1 (usually employment)	16.01
burden_interact	SU status time × burden	7.61
burden	Number of items reporting	7.07
major_producer_flag1	Dummy variable, 1 = yes, considered a major producer based on size	2.99
season	Calendar season (e.g. winter=4, spring = 3,...)	1.19
ipp_ppi_overlap1	dummy variable, 1 = yes if providing data for MXPI and PPI	0.3
su_status_time	Cumulative time in status, months	0
Note: MXPI is the International Price Program at BLS, which also uses voluntarily provided industrial survey data. No overlap possible between CPI and PPI since CPI is not an industrial price survey.		

None of the features used in the model were especially correlated, indicating no risk of multicollinearity—though Random Forest models are typically robust against the presence of multicollinearity. The data, once broken down into two classes (return the following month and not return), showed a strong class imbalance: 91 percent of the observations were labeled "no" (as in they did not return to full response status the following month) and only 9 percent was "yes" (they did return to full response status). To address this issue and prevent the classifier from being biased toward "no", a weighted cost approach was employed, whereby the classifier was penalized with more errors for misclassifying the minority class ("Yes") to correct bias toward the majority class. These weighted costs are incorporated via a cost matrix, shown in Table 6 below. Additionally, the Random Forest was trained using case weights equal to inverse class frequency. This method was employed to give the minority class (the "yes") greater weight then the majority class ("no") when training the model.

	Table 6: Final Cost Matrix		
		Actual	
		No	**Yes**
Predicted	**No**	0	5
	Yes	30	0

The variables/features included in the model were chosen via repeated 10-fold cross-validation, repeated five times: This was done to ensure the model results were not simply "one-off" or due to a single, random sample. It was also done to ensure the best combination of (available) variables was included in

the model, listed in Table 5 above. As Table 5 shows, the strongest predictor of whether an SU will move from full nonresponse to full response is the cumulative number of times the SU has made the named transition between states, followed by total time in PPI, then a measure of SU size, which is typically the total employment at the given establishment. The interaction term of respondent burden and cumulative time in status is shown to have a moderate impact, but less than total SU weight. This suggests that long-term respondents with larger index weight are more likely to move from full nonresponse to full response.

Table 7 below shows the confusion matrix for the final model. As Table 7 shows, the model performs reasonably well. The overall accuracy rate of 0.95 is good but heavily influenced by the heavily imbalanced dataset. The P-value calculated for McNemar's Test indicates the classifier is good at distinguishing cases in which a respondent moved from full nonresponse to full response the next month ("yes") and cases in which they didn't ("no"). This is partly due to the decision threshold used. The standard threshold for binary classification is 0.5. That is, if the probability of positive classification is above 0.5 the prediction is "yes" (as in an SU will move from full nonresponse status to full response status next month). The initial threshold employed was conservative—0.8, the value at which Youden's J statistic was maximized. This was done to minimize false positives. A false positive would indicate that the model predicted an SU would return from nonresponse, but they failed to. A false negative would be the opposite. Since the goal is to learn which respondents should be contacted, a false positive would amount to a waste of valuable resources—time. A false negative in this context is harmless.

By implementing the cost matrix outlined in Table 6, the classifier is penalized for false negatives as well as false positives but significantly more so for false positives. The confusion matrix resulting from the use of these hyperparameters is presented below in Table 7.

Table 7: Confusion Matrix & Metrics, Test Data—Final Run

		Actual Outcome	
		No	**Yes**
Predicted Outcome	**No**	10774	452
	Yes	108	597
	Overall Accuracy	0.95	
	Mcnemar's Test P-value	0.00	
	Precision	**0.85**	
	Recall (TPR)	**0.57**	
	False Positive Rate	**0.01**	
	True Negative Rate	**0.99**	
	Balanced Accuracy	**0.78**	

As Table 7 shows, the cost matrix was employed to good effect. The Precision score of 0.85 indicates that 85 percent of "yes" predictions were correct. Combined with the especially low false positive rate of approximately 1 percent, the output indicates the model could be effectively employed in a production context. Overall, what these results suggest is that supervised Machine Learning tools can direct which resources should be used to convert nonrespondents into full respondents.

6. Conclusion

The purpose of this paper was to provide a more robust understanding of PPI response and nonresponse patterns by taking an aggregate, holistic approach to data analysis. By aggregating individual item-level data, a complete history has been constructed for each respondent initiated into PPI over a ten-year period. Each respondent was tracked for up to 84 months, or the equivalent of a standard PPI sample.

Because the data does not follow a specific process throughout the repricing history of each individual respondent, it is treated as a stochastic process, which allows the estimation of probabilities for each individual repricing state. The data shows that in the aggregate, respondents are most likely to provide a full suite of data for all items requested by the PPI. The probability that a given respondent continues to provide all the data requested, predictably declines with their time in the system. As the probability of full response declines, the likelihood of full nonresponse rises—to a point.

After about 25 months in the PPI survey, the probability of a fully productive respondent becoming fully nonproductive flattens to about 11 percent. This indicates that any activity aimed at curbing nonresponse would be most effective before the 25th month in the system. More specifically, based on graphical evidence in Figure 2, the point at which activities to curb nonresponse begin to become less effective is at about 8 months in the PPI survey. This is because the slope of the curve depicting the probability of staying in full nonresponse status is steepest prior to month 8, indicating the largest changes. After month 8, the typical respondent is more likely to be in full response status than nonresponse status, as shown by the vertical distance between the blue and red curves. Furthermore, the slope of the blue curve softens, indicating a substantial change in the rate at which nonrespondents are converted into fully productive ones. This continues until about the 25th month in the survey, after which both the full response and nonresponse curves flatten, indicating that inertia has essentially taken over—after this point there are no significant changes in either probability.

The analysis above shows that BLS economists and statisticians may address the issue of nonresponse in the PPI - and perhaps for its other economic statistics that rely on surveys - through implementing supervised Machine Learning tools. By feeding survey data used by the Markov model into a well-tuned Random Forest mode with task-specific hyperparameters, BLS can determine and assign probabilities to respondents who are most likely to move from nonresponse to full response. BLS economists and statisticians are currently considering implementation options most appropriate for the large-scale, monthly PPI.

Acknowledgements

Kathleen Frawley, Melanie Santiago, Susan Fleck, Daniel Yang, Steve Paben all provided valuable comments.

References

Cox, D.R. and Miller, H.D. 1987. The Theory of Stochastic Processes. New York: Chapman & Hall.

Medhi, J. 2003. *Stochastic Models in Queueing Theory*. San Diego: Elsevier Science.

U.S. Dept. of Labor, Bureau of Labor Statistics. 2016. BLS Handbook of Methods. Chapter 14, "Producer Prices." Washington, DC. Retrieved from: https://www.bls.gov/opub/hom/pdf/homch14.p