PREDICTION WITH INCOMPLETE INFORMATION

THIS VERSION: April 7, 2025

Megan Czasonis, Mark Kritzman, and David Turkington

Megan Czasonis is a managing director at State Street Associates in Cambridge, MA.
mczasonis@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Mark Kritzman is the chief executive officer of Windham Capital Management in Cambridge, MA, and a senior lecturer at the MIT Sloan School of Management in Cambridge, MA.
kritzman@mit.edu
100 Main Street, Cambridge MA, 02142

David Turkington is senior managing director and head of State Street Associates in Cambridge, MA.
dturkington@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Key Takeaways

The simplest way to treat incomplete information is to use only predictive variables without missing information, but this approach is unappealing because it discards valuable information.

Alternatively, one may use statistical techniques to manufacture missing information thereby preserving available information, but these techniques often depend on limiting assumptions and produce tenuous results.

A new prediction technique called relevance-based prediction (RBP) treats observations with missing information in a way that preserves remaining information and explicitly accounts for the relative importance of observations with missing information when forming a prediction and assessing its reliability.

Abstract

A key requirement for forming data driven predictions is to assemble the best possible set of observations for the predictive variables. This task, however, is not often easy. In the case of time series data, some variables have shorter histories than others, and some variables are reported less frequently than others. And in the case of cross-sectional data, some information is not reported for every case. We are therefore faced with several choices. We can discard predictive variables with missing information. We can exclude observations with missing information and retain only those with full information for all the predictive variables. We can use statistical techniques to manufacture replacements for the missing information. However, each of these approaches has significant drawbacks. The authors propose a new procedure for treating missing information that enables us to retain as much information as possible to form predictions and, at the same time, to account for the relative reliability of observations with missing information.

PREDICTION WITH INCOMPLETE INFORMATION

When we set out to form data driven predictions, we first identify predictive variables that we believe to be important to the prediction, and then we assemble a sample of observations for those predictive variables. It is often the case, though, that some observations do not have complete information. For example, if we form our prediction from a time series of observations, some predictive variables may have shorter histories than others or some observations may be reported less frequently than others. And if we form our prediction from cross sectional data, some observations may have missing information. We could discard predictive variables that have missing information, or we could use all the predictive variables but only include observations with full information that are common across all predictive variables. Both these approaches are unappealing, though, because they force us to discard useful information. Alternatively, we could use statistical techniques to manufacture replacement information, but these techniques rely on limiting assumptions and often yield tenuous results.[1] Moreover, these techniques fail to distinguish between the relative importance of missing information in one prediction task compared to another. We propose a fourth alternative which follows from a new prediction technique called relevance-based prediction (RBP). This technique, which has been shown to extract as much information from complex datasets as machine learning models,[2] offers an elegant solution for treating observations with missing information that allows us to retain as much information as possible and, at the same time, account for the relative reliability of the observations with missing information.

We proceed as follows. We first describe RBP and its key features: relevance, fit, and grid prediction. We then describe a simple but powerful way to treat missing information within the context of RBP, which is to assign a zero to a missing observation's relevance weights within a grid that considers every possible combination of variables. We present a simple illustration of how this approach retains observations that eliminating observations or omitting a variable would discard. Next, we explain how RBP's treatment of missing information accounts for the relative importance of missing information, and we present a toy example for a single grid cell prediction to support our argument. Following that we present more toy examples to illustrate how the assignment of zero to the relevance weight of observations with missing information affects the composite prediction that flows from the prediction grid. We then carry out a simulation to provide further evidence that RBP's approach to treating missing information reliably captures the relative importance of the missing information. We conclude with a summary.

Relevance-Based Prediction

Relevance-based prediction (RBP) is a model-free prediction routine that forms a prediction as a weighted average of observed outcomes in which the weights are based on a precisely defined statistic called relevance. RBP also relies on fit which measures the reliability of each individual prediction, and grid prediction which gives a composite prediction from many predictions formed from different combinations of observations and predictive variables.[3] As we will show, the grid's composite prediction explicitly accounts for the relative reliability of missing observations by the way it is formed.

Relevance

Relevance is a precise statistical measure of the importance of an observation to a prediction. It is composed of similarity and informativeness, which are both measured as Mahalanobis distances, as shown by Equations 1 through 4.[4]

$$r_{it} = sim(x_i, x_t) + \tfrac{1}{2}\big(info(x_i, \bar{x}) + info(x_t, \bar{x})\big) \tag{1}$$

$$sim(x_i, x_t) = -\tfrac{1}{2}(x_i - x_t)\Omega^{-1}(x_i - x_t)' \tag{2}$$

$$info(x_i, \bar{x}) = (x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \tag{3}$$

$$info(x_t, \bar{x}) = (x_t - \bar{x})\Omega^{-1}(x_t - \bar{x})' \tag{4}$$

In Equations 1 through 4, x_i is a vector of the values of K predictive variables for a prior observation, x_t is a vector of the values of the predictive variables for a specific prediction task, $\bar{x} = 1_N 1_N' X N^{-1}$ is the average of the predictive variables across all observations, and Ω^{-1} is the inverse covariance matrix of all the observations of the variables. The vector $(x_i - x_t)$ measures how distant each variable's observed value is from its corresponding value in the prediction task, when measured in isolation. By multiplying this vector by the inverse covariance matrix, we capture the interaction of the predictive variables, and at the same time we standardize the distances by dividing by variance. By multiplying this product by the transpose of the vector $(x_i - x_t)$ we consolidate the outcome into a single number. All else being equal, observations that are like current circumstances but different from average circumstances are more relevant than those that are not.

This definition of relevance is not arbitrary. We know from information theory that the information contained in an observation is the negative logarithm of its likelihood.[5] We also know from the Central Limit Theorem that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance. Therefore, the information contained in a point on a multivariate normal distribution is proportional to a Mahalanobis distance.

Relevance-based prediction forms a prediction as a weighted average of prior outcomes for Y.

$$\hat{y}_t = \sum_{i=1}^{N} w_{it} y_i \tag{5}$$

If we define weights in terms of relevance as follows, which admits the relevance-weighted average of every prior outcome in the observed data sample, the result is precisely equivalent to the prediction that results from linear regression analysis.[6]

$$w_{it,linear} = \frac{1}{N} + \frac{1}{N-1} r_{it} \tag{6}$$

In most cases, however, we can produce a more reliable prediction by censoring the observations that are less relevant than a chosen threshold, which leads to the following definition of prediction weights.

$$w_{it,retained} = \frac{1}{N} + \frac{\lambda^2}{n-1}\left(\delta(r_{it}) r_{it} - \varphi \bar{r}_{sub}\right) \tag{7}$$

$$\delta(r_{it}) = \begin{cases} 1 & if\ r_{it} \geq r^* \\ 0 & if\ r_{it} < r^* \end{cases} \tag{8}$$

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,retained}} = \frac{\frac{1}{N-1}\sum_{i=1}^{N} r_{it}^2}{\frac{1}{n-1}\sum_{i=1}^{N} \delta(r_{it}) r_{it}^2} \tag{9}$$

In Equations 6 through 9, $n = \sum_{i=1}^{N} \delta(r_{it})$ is the number of observations that are fully retained, $\varphi = n/N$ is the fraction of observations in the retained sample, and $\bar{r}_{sub} = \frac{1}{n}\sum_{i=1}^{N} \delta(r_{it}) r_{it}$ is the average relevance value of the observations in the retained sample. It is important to note that $w_{it,retained}$ depends crucially on the prediction circumstances x_t. Relevance is reassessed for each prediction circumstance which further affects the identification of the retained subsample and introduces nonlinear conditional dependence of the prediction $\hat{y}_t$ on the prediction circumstances x_t. The scaling factor λ^2 compensates for a bias that would otherwise result from relying on a small subsample of highly relevant observations. In the case of linear regression analysis $n = N$ and $\lambda^2 = 1$. Lastly, note that the regression weights always sum to 1.[7]

Fit

Fit quantifies the prevalence of useful patterns in a dataset, which provides a principled way to evaluate the relative efficacy of alternative calibrations for each prediction task. Additionally, fit reveals how much confidence we should have in a specific prediction task, separately from the confidence we have in the overall prediction routine.

Consider a pair of observations that are used to form a prediction. Each observation has a weight and an outcome. We are interested in the alignment of the weights of the two observations with their outcomes. We first standardize them by subtracting the average value and dividing this difference by standard deviation – in essence, converting them to z-scores. We then measure their alignment by taking the product of these standardized values. If the product is positive, their relevance is aligned with their outcomes, and the larger the product, the stronger the alignment. We perform this calculation for every pair of observations in our sample. We should also note that all the formulas we have thus far considered for weights rely only on relevance, which in turn relies only on the x_is, the x_t, and the $\bar{x}$. They do not use any of the information from observed outcomes. To determine fit, however, we must consider outcomes (the y_is).

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j z_{w_{it}} z_{w_{jt}} z_{y_i} z_{y_j} \qquad (10)$$

Equation 11 intuitively describes fit as the squared correlation of relevance weights and outcomes, which conceptually matches the notion of the conventional R-squared statistic. As we soon show, this connection of fit to R-squared is critically important.

$$fit_t = \rho(w_t, y)^2 \qquad (11)$$

Although we compute fit from the full sample of observations, the weights that determine fit vary with the threshold we choose to define the relevant subsample. As we focus the subsample on observations that are more relevant, we should expect the fit of the subsample to increase, but we should also expect more noise as we shrink the number of observations. The fit across pairs of all observations in the full sample implicitly captures this tradeoff between subsample fit and noise by overweighting observations that are more relevant and underweighting observations that are less relevant.

Like relevance, fit is not arbitrary. In the case of linear regression analysis with $n = N$, the informativeness-weighted average fit across all prediction tasks in the observed sample equals R-squared.[8]

$$R^2 = \frac{1}{N-1}\sum_{t=1}^{N} info(x_t, \bar{x})fit_t \tag{12}$$

Censoring observations that fall below a relevance threshold is more effective to the extent there is asymmetry between the fit of the weights formed from the retained subsample of observations and the fit of the weights formed from the complementary set of censored observations. We measure asymmetry between the fit of the retained and censored subsamples as shown by Equation 13. The $(+)$ superscript designates weights formed from the retained observations while the $(-)$ superscript designates weights formed from the censored observations. Asymmetry recognizes the benefit of censoring non-relevant observations that contradict the predictive relationships that exist among the relevant observations. This assessment also inherently considers the relative sample sizes of the two subsamples.

$$asymmetry_t = \frac{1}{2}\left(\rho\left(w_t^{(+)}, y\right) - \rho\left(w_t^{(-)}, y\right)\right)^2 \tag{13}$$

To calculate adjusted fit, we add asymmetry to fit and multiply this sum by K, the number of predictive variables included in the prediction, as shown by Equation 14. Multiplication by the number of predictive variables allows us to compare predictions based on different numbers of predictive variables. Adjusted fit recognizes that we are more likely to observe a spurious relationship from prediction weights based on just one or a few variables than we are based on a collection of many variables.

$$adjusted\ fit_t = K(fit_t + asymmetry_t) \tag{14}$$

Grid Prediction

Grid prediction employs a grid in which the columns represent different combinations of predictive variables, and the rows represent subsamples of observations determined by different relevance thresholds. Each cell contains a prediction and an associated adjusted fit. The assessment of reliability using adjusted fit occurs before the prediction is rendered and the subsequent outcome is known. Grid prediction forms a composite prediction as a reliability-weighted average of the predictions from all possible calibrations. Equation 15 defines reliability weights, ψ_θ, as the adjusted fit for a parameter calibration, θ, divided by the sum of all adjusted fits across all parameter calibrations.

$$\psi_\theta = \frac{adjusted\ fit_\theta}{\sum_{\tilde{\theta}} adjusted\ fit_{\tilde{\theta}}} \tag{15}$$

Equation 16 describes the composite prediction.

$$\hat{y}_{t,grid} = \sum_\theta \psi_\theta \hat{y}_{t,\theta} \tag{16}$$

Exhibits 1 and 2 illustrate how RBP forms a prediction. Exhibit 1 shows how we compute the prediction for a single cell in the prediction grid. It includes hypothetical values for the X and Y variables. The panel on the right gives values for the similarity and informativeness of prior observations and the informativeness of the observations for the current prediction task. It also shows the relevance of each prior observation and the observation's relevance weight.

Exhibit 1: Single Cell Prediction

Variables	Y	X_1	X_2	X_3	X_4	X_5	X_6	Similarity	Info$_i$	Info$_t$	Relevance	Weight
Prediction Task t	?	2.78	8.75	0.28	0.61	0.31	0.58					
Observation 1	20.67	3.13	10.21	0.29	0.00	0.47	0.53	-4.30	12.13	11.96	7.75	4.9%
Observation 2	6.30	4.14	12.24	0.21	0.60	0.29	0.48	-4.06	2.99	11.96	3.41	2.0%
Observation 3	5.19	1.99	9.78	0.16	0.52	0.10	0.48	-7.36	2.43	11.96	-0.17	-0.4%
Observation 4	10.41	3.21	13.47	0.26	0.34	0.48	0.54	-3.41	3.94	11.96	4.54	2.7%
•	•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•	•
Observation n	4.49	4.14	3.14	0.23	0.31	0.22	0.37	-7.36	2.75	11.96	-0.01	-0.4%
			Prediction						19.40			
			Adjusted Fit:						2.32			

Exhibit 2 gives a visual representation of grid prediction. The columns represent different subsets of variables, and the rows represent different subsamples of observations as determined by different relevance thresholds. Each cell represents a calibration θ; that is, a unique combination of predictive variables and observations. In practice, we would consider all 31 combinations of five variables, but for illustrative purposes we show only seven columns in Exhibit 2. The first values shown in the cells are the calibration-specific predictions $\hat{y}_t$ for a given prediction task t. The second values are the weights ψ_θ we apply to the calibration-specific predictions to form the composite prediction. The values in the grid are specific to each prediction task. This illustration gives a composite prediction of 16.30 (15.7 x 1.72% + 15.7 x 1.15% + 10.1 x 0.24% + . . . + 9.3 x 0.04%).

Exhibit 2: Grid Prediction – Illustrative Example

	Variable Combinations													
Observation Censoring Threshold	$X_1 X_2 X_3 X_4 X_5 X_6$		$X_1 X_2 X_3 X_4$		$X_1 X_3 X_4$		$X_2 X_5 X_6$		$X_3 X_6$		X_2		X_6	
0.0	15.7	1.72%	15.7	1.15%	10.1	0.24%	15.3	1.37%	10.9	0.54%	15.3	0.47%	7.4	0.06%
0.1	16.4	2.02%	16.7	1.39%	10.4	0.23%	15.4	1.88%	12.5	0.73%	15.5	0.50%	7.7	0.04%
0.2	17.5	2.20%	17.4	1.43%	10.3	0.18%	15.4	1.91%	12.6	0.64%	15.5	0.44%	7.9	0.05%
0.3	17.8	2.17%	17.7	1.43%	10.5	0.20%	15.5	2.24%	12.6	0.62%	15.5	0.42%	7.9	0.05%
0.4	18.2	2.29%	18.0	1.50%	10.6	0.22%	15.4	2.18%	12.7	0.65%	15.5	0.41%	8.1	0.07%
0.5	18.6	2.50%	18.2	1.58%	10.7	0.25%	14.3	2.50%	12.8	0.70%	15.3	0.41%	8.1	0.06%
0.6	18.7	2.47%	18.4	1.61%	10.7	0.23%	15.4	1.21%	13.1	0.73%	15.4	0.42%	8.8	0.10%
0.7	19.0	2.47%	18.8	1.63%	10.7	0.19%	15.4	2.20%	12.9	0.62%	15.4	0.41%	8.7	0.07%
0.8	**19.4**	**2.32%**	19.1	1.50%	11.5	0.20%	15.3	2.04%	13.7	0.57%	15.5	0.37%	8.6	0.04%
0.9	19.5	1.26%	18.8	0.81%	12.9	0.22%	15.5	1.73%	14.0	0.32%	15.3	0.25%	9.3	0.04%

Composite Prediction : 16.30

Note that each cell's prediction is a linear function of observations, and the grid prediction is a linear function of each cell's prediction. Therefore, we can express the grid prediction in terms of composite weights applied to each observation, as shown by Equation 17. Composite weights are important because they preserve the transparency of each observation's contribution to the current prediction task, and they allow us to calculate fit from composite weights as a final gauge of the grid prediction's reliability.

$$w_{it,grid} = \sum_{\theta} \psi_{\theta} w_{it,\theta} \tag{17}$$

RBP and Missing Data

RBP treats missing data by assigning a relevance weight of zero to observations with missing information, and aggregates alternative uses of data across many variable combinations in a prediction grid. This approach has two important virtues. First, it allows us to retain more observations than if we were to eliminate the observations with missing information or if we were to omit predictive variables with missing information. Second, it allows us to account for the relative importance of observations with missing information. We first discuss how RBP preserves information.

Imagine we are presented with the data shown in Exhibit 3, in which three values are missing for variable X1 and two values are missing for variable X2.

Exhibit 3: Potential Samples for Analysis

	X1 and X2		X1 only		X2 only	
	2 Variables 5 Observations		1 Variable 7 Observations		1 Variable 8 Observations	
	X1	X2	X1	X2	X1	X2
1	0.35	1.11	0.35	1.11	0.35	1.11
2		-0.52		-0.52		-0.52
3		-1.60		-1.60		-1.60
4		-0.22		-0.22		-0.22
5	0.12	-0.90	0.12	-0.90	0.12	-0.90
6	-0.37	0.17	-0.37	0.17	-0.37	0.17
7	0.80		0.80		0.80	
8	-0.26		-0.26		-0.26	
9	0.60	0.34	0.60	0.34	0.60	0.34
10	0.17	-0.25	0.17	-0.25	0.17	-0.25

The three panels in Exhibit 3 illustrate alternative approaches to extracting subsets of data that are amenable to traditional analysis. In the left panel, we remove all observations that contain missing data, preserving five observations of both X1 and X2 but ignoring five pieces of information. In the middle panel, we remove variable X2 entirely, preserving seven observations of variable X1 but ignoring eight pieces of information. And in the right panel, we remove variable X1 entirely, preserving eight observations of variable X2 but ignoring seven pieces of information. Viewed in isolation, each of these approaches is suboptimal because it sacrifices potentially useful information.

RBP addresses this issue by blending the information from each panel using the inherent properties of grid prediction. In this simplified example, the grid consists of three cells. Recall that RBP forms a prediction by taking a weighted average of observed outcomes in which the weights are based on relevance. For the first cell, we assign relevance weights to each of the available observations and set the remaining observation weights to zero. Then we compute the adjusted fit of this cell's prediction weights. It is crucial to recognize that adjusted fit is penalized for the fact that five observations have zero weights. Intuitively, this occurs because fit equals the squared correlation between weights and outcomes (equation 11), and zero weights dilute this correlation. For the second cell, we ignore variable X2 completely, assign relevance weights to the set of seven available observations based on X1, and set the remaining weights to zero. As before, we compute the adjusted fit of this cell's prediction weights. The adjusted fit of this cell is only dampened by three zero weights, which is fewer than the previous cell. However, this cell's adjusted fit is dampened by having only one variable, K, rather than two. A similar situation applies to the third cell.

The presence of missing data leads to a fundamental tradeoff: cells with more variables tend to have fewer observations. Adjusted fit accounts for this tradeoff while measuring the precise predictive value of the information in each cell. From Equation 17, the final weight of each observation is a blended average across cells. Thus, all 10 observations in this example will receive a nonzero weight that reflects both the availability of data for that observation and the efficacy of the data for cells where it is available. This blending relies on fit, which in turn relies on relevance. If we were to instead employ a model-based approach to prediction such as a linear regression or a neural network, we would not be able to implement this approach because we would not know the impact of each observation on the prediction. Instead, we would be forced to choose among suboptimal reductions of the information set.

Now let us consider how RBP's treatment of missing information accounts for the relative importance of observations across different prediction tasks. If the prediction in a cell in the prediction grid is based, in part, on unimportant missing information, the prediction will not differ meaningfully from a prediction that included unimportant information because the relevance weight of the observation would be close to zero anyway. Also, the reliability of a cell's prediction that is based on missing unimportant information would not differ much from the reliability that would obtain if the unimportant missing information were included in the cell's formation of the prediction. Therefore, assigning zero to observations with missing unimportant information has a minimal impact on the grid's composite prediction, because it is based on the relative reliability of each cell's prediction. The opposite is true for observations that are missing important information. The prediction of a cell that is missing important information will change meaningfully from the prediction that would occur if the important information were included, as would the cell's reliability weight in the grid's composite prediction. These effects show how assigning zero to observations with missing information automatically accounts for the relative importance of observations with missing information. We illustrate these effects in Exhibits 4 through 7.

The Effect of Missing Information on Single Cell Prediction

In Exhibit 4, the panels on the left show the calculation of fit for a single cell prediction in the prediction grid based on a sample of observations with complete information. We consider a sample of only four observations for the sake of transparency. These results follow from Equation 10 which gives the average alignment between the standardized values of relevance weights and outcomes for all pairs of observations that go into a prediction task. The middle panels show the same calculation of fit, but with an observation that has relatively unimportant missing information, and which is given a relevance weight of zero. Notice that fit changes only slightly from 0.75 to 0.71. The panels on the right again show the same calculation of fit, but this time with an observation that has relatively important missing information, and which is given a relevance weight of zero. In this case fit changes significantly from 0.75 to 0.25.

Exhibit 4: The Effect of Missing Information on the Fit of a Single Cell Prediction

Complete Sample

Observation	w	y	z(w)	z(y)
1	20%	3.16	-0.39	0.20
2	30%	1.87	0.39	-0.30
3	40%	5.91	1.16	1.25
4	10%	-0.34	-1.16	-1.15
Average	25%	2.65	0.00	0.00
Standard Deviation	13%	2.61	1.00	1.00

Missing Unimportant Information

Observation	w	y	z(w)	z(y)
1	20%	3.16	-0.26	0.20
2	35%	1.87	0.51	-0.30
3	45%	5.91	1.02	1.25
4	0%	-0.34	-1.28	-1.15
Average	25%	2.65	0.00	0.00
Standard Deviation	20%	2.61	1.00	1.00

Missing Important Information

Observation	w	y	z(w)	z(y)
1	35%	3.16	0.51	0.20
2	45%	1.87	1.02	-0.30
3	0%	5.91	-1.28	1.25
4	20%	-0.34	-0.26	-1.15
Average	25%	2.65	0.00	0.00
Standard devia	20%	2.61	1.00	1.00

Pairs						
A	B	$z(w_i)$	$z(y_i)$	$z(w_j)$	$z(y_j)$	Product
1	1	-0.39	0.20	-0.39	0.20	0.01
1	2	-0.39	0.20	0.39	-0.30	0.01
1	3	-0.39	0.20	1.16	1.25	-0.11
1	4	-0.39	0.20	-1.16	-1.15	-0.10
2	1	0.39	-0.30	-0.39	0.20	0.01
2	2	0.39	-0.30	0.39	-0.30	0.01
2	3	0.39	-0.30	1.16	1.25	-0.17
2	4	0.39	-0.30	-1.16	-1.15	-0.15
3	1	1.16	1.25	-0.39	0.20	-0.11
3	2	1.16	1.25	0.39	-0.30	-0.17
3	3	1.16	1.25	1.16	1.25	2.11
3	4	1.16	1.25	-1.16	-1.15	1.93
4	1	-1.16	-1.15	-0.39	0.20	-0.10
4	2	-1.16	-1.15	0.39	-0.30	-0.15
4	3	-1.16	-1.15	1.16	1.25	1.93
4	4	-1.16	-1.15	-1.16	-1.15	1.77
		Sum:			6.71	
		$(N-1)^2$:			9	
		Fit:			0.75	

Pairs						
A	B	$z(w_i)$	$z(y_i)$	$z(w_j)$	$z(y_j)$	Product
1	1	-0.26	0.20	-0.26	0.20	0.00
1	2	-0.26	0.20	0.51	-0.30	0.01
1	3	-0.26	0.20	1.02	1.25	-0.06
1	4	-0.26	0.20	-1.28	-1.15	-0.07
2	1	0.51	-0.30	-0.26	0.20	0.01
2	2	0.51	-0.30	0.51	-0.30	0.02
2	3	0.51	-0.30	1.02	1.25	-0.19
2	4	0.51	-0.30	-1.28	-1.15	-0.22
3	1	1.02	1.25	-0.26	0.20	-0.06
3	2	1.02	1.25	0.51	-0.30	-0.19
3	3	1.02	1.25	1.02	1.25	1.63
3	4	1.02	1.25	-1.28	-1.15	1.87
4	1	-1.28	-1.15	-0.26	0.20	-0.07
4	2	-1.28	-1.15	0.51	-0.30	-0.22
4	3	-1.28	-1.15	1.02	1.25	1.87
4	4	-1.28	-1.15	-1.28	-1.15	2.14
		Sum:			6.43	
		$(N-1)^2$:			9	
		Fit:			0.71	

Pairs						
A	B	$z(w_i)$	$z(y_i)$	$z(w_j)$	$z(y_j)$	Product
1	1	0.51	0.20	0.51	0.20	0.01
1	2	0.51	0.20	1.02	-0.30	-0.03
1	3	0.51	0.20	-1.28	1.25	-0.16
1	4	0.51	0.20	-0.26	-1.15	0.03
2	1	1.02	-0.30	0.51	0.20	-0.03
2	2	1.02	-0.30	1.02	-0.30	0.09
2	3	1.02	-0.30	-1.28	1.25	0.49
2	4	1.02	-0.30	-0.26	-1.15	-0.09
3	1	-1.28	1.25	0.51	0.20	-0.16
3	2	-1.28	1.25	1.02	-0.30	0.49
3	3	-1.28	1.25	-1.28	1.25	2.54
3	4	-1.28	1.25	-0.26	-1.15	-0.47
4	1	-0.26	-1.15	0.51	0.20	0.03
4	2	-0.26	-1.15	1.02	-0.30	-0.09
4	3	-0.26	-1.15	-1.28	1.25	-0.47
4	4	-0.26	-1.15	-0.26	-1.15	0.09
		Sum:			2.27	
		$(N-1)^2$:			9	
		Fit:			0.25	

The Effect of Missing Information on Grid Prediction

We next extend our analysis to show how assigning zero to the relevance weights of observations with missing information affects the composite prediction that comes from the prediction grid.

Exhibit 5 depicts a grid prediction assuming there is complete information. There are three combinations of predictive variables: X1 and X2, X1, and X2. And there are two subsamples of observations; one with a relevance threshold of 0, which means that the full sample of observations is used, and one with a relevance threshold of 0.5, which means that the 50% most relevant observations are used. The prediction for the calibration that uses all the predictive variables and all the observations equals 2.52. It is calculated by multiplying the Y outcomes by the relevance weights in the column labeled r* = 0 under the broader column heading (X1 and X2) and summing these products. The grid cell weights are calculated as the relative adjusted fits as given by Equation 15. The composite prediction 2.19 is calculated by summing the products of the Y outcomes and grid cell weights across all the calibrations.

It might be helpful to reconcile the format of this exhibit with the prediction grid shown in Exhibit 2. Given this dataset, the prediction grid has three columns and two rows. The prediction 2.52 and its associated adjusted fit of 34.9% would be the values in the upper left cell of the prediction grid. The prediction 2.38 and its adjusted fit of 22.8% would go in the cell in the first column and second row of the prediction grid. The prediction 1.81 with adjusted fit of 14.7% would go into the cell in the first row and second column of the prediction and so on.

Exhibit 5: Grid Prediction with Complete Information

N	Inputs			Prediction Weights						Grid
	X1	X2	Y	X1 and X2		X1		X2		
				r* = 0	r* = 0.5	r* = 0	r* = 0.5	r* = 0	r* = 0.5	
1	0.35	1.11	1.16	24.3%	28.7%	13.4%	9.1%	26.0%	30.6%	23.3%
2	-1.79	-0.52	-3.82	-5.2%	-0.8%	-8.9%	1.8%	3.6%	1.9%	-2.6%
3	-1.21	-1.60	-3.06	-13.4%	-0.8%	-2.8%	1.8%	-11.2%	1.9%	-6.4%
4	1.83	-0.22	1.29	18.4%	16.5%	28.8%	42.5%	7.8%	-2.0%	18.3%
5	0.12	-0.90	-0.64	1.6%	-0.8%	11.0%	1.8%	-1.5%	1.9%	2.1%
6	-0.37	0.17	0.04	10.2%	-0.3%	5.9%	1.8%	13.1%	7.6%	6.7%
7	0.80	1.59	3.13	31.9%	44.6%	18.1%	19.4%	32.6%	42.4%	32.8%
8	-0.26	-0.29	-0.33	5.9%	-0.8%	7.1%	1.8%	6.8%	1.9%	4.0%
9	0.60	0.34	2.55	17.5%	14.7%	16.0%	14.9%	15.5%	11.7%	15.7%
10	0.17	-0.25	-0.25	8.8%	-0.8%	11.5%	5.1%	7.3%	1.9%	6.0%
Prediction				2.52	2.38	1.81	1.49	2.00	1.80	2.19
Fit				0.86	0.56	0.73	0.33	0.62	0.38	0.80
Asymmetry				0.00	0.01	0.00	0.04	0.00	0.01	
Adjusted Fit				1.72	1.13	0.73	0.36	0.62	0.38	
Adjusted Fit Weight				34.9%	22.8%	14.7%	7.3%	12.6%	7.7%	

Exhibit 6 gives the same information as Exhibit 5 for a case in which we remove relatively unimportant information from two of the observations. It is worth noting that the composite prediction is (approximately) the same as the prediction produced from the sample with complete information and that the fit only declines by a small amount.

Exhibit 6: Grid Prediction with Unimportant Missing Information

N	Inputs			Prediction Weights						Grid
	X1	X2	Y	X1 and X2		X1		X2		
				r* = 0	r* = 0.5	r* = 0	r* = 0.5	r* = 0	r* = 0.5	
1	0.35	1.11	1.16	24.6%	28.2%	14.6%	9.8%	26.5%	30.5%	23.6%
2	-1.79	-0.52	-3.82	-4.0%	-0.9%	-7.9%	2.3%	4.5%	1.8%	-1.9%
3	-1.21	-1.60	-3.06	-11.6%	-0.9%	-1.8%	2.3%	-10.1%	1.8%	-5.3%
4	1.83	-0.22	1.29	19.7%	16.3%	30.2%	43.2%	8.6%	1.8%	19.4%
5		-0.90	-0.64	0.0%	0.0%	0.0%	0.0%	-0.6%	1.8%	0.1%
6	-0.37	0.17	0.04	11.1%	-0.9%	7.1%	2.3%	13.9%	6.9%	7.1%
7	0.80	1.59	3.13	32.0%	46.1%	19.4%	20.1%	33.0%	42.6%	33.5%
8	-0.26		-0.33	0.0%	0.0%	8.2%	2.3%	0.0%	0.0%	1.4%
9	0.60	0.34	2.55	18.3%	13.1%	17.3%	15.6%	16.1%	11.1%	15.9%
10	0.17	-0.25	-0.25	9.9%	-0.9%	12.8%	2.3%	8.1%	1.8%	6.4%
Prediction				2.49	2.37	1.91	1.53	1.99	1.85	2.19
Fit				0.82	0.54	0.72	0.32	0.60	0.40	0.76
Asymmetry				0.00	0.01	0.00	0.04	0.00	0.00	
Adjusted Fit				1.64	1.10	0.72	0.37	0.60	0.41	
Adjusted Fit Weight				34.0%	22.8%	14.8%	7.6%	12.3%	8.4%	

Exhibit 7 shows the results of grid prediction for a case in which we remove relatively important information from two of the observations. In this case, the prediction changes significantly from the case in which there is complete information, and the adjusted fit of this prediction declines sharply.

Exhibit 7: Grid Prediction with Important Missing Information

N	Inputs			Prediction Weights						Grid
	X1	X2	Y	X1 and X2		X1		X2		
				r* = 0	r* = 0.5	r* = 0	r* = 0.5	r* = 0	r* = 0.5	
1	0.35	1.11	1.16	54.1%	79.0%	15.6%	11.6%	46.8%	69.1%	42.4%
2	-1.79	-0.52	-3.82	-3.4%	-5.7%	-9.1%	2.4%	3.6%	-2.9%	-3.2%
3	-1.21	-1.60	-3.06	-29.0%	-5.7%	-2.4%	2.4%	-24.9%	-2.9%	-12.2%
4	1.83	-0.22	1.29	27.2%	24.2%	32.7%	46.6%	11.7%	-1.8%	25.5%
5	0.12	-0.90	-0.64	-1.6%	-5.7%	12.9%	2.4%	-6.2%	-2.9%	1.1%
6	-0.37	0.17	0.04	24.2%	18.2%	7.3%	2.4%	22.0%	19.1%	15.5%
7	0.80		3.13	0.0%	0.0%	20.9%	22.4%	0.0%	0.0%	8.1%
8	-0.26	-0.29	-0.33	12.4%	-5.7%	8.6%	2.4%	9.8%	-2.9%	5.9%
9		0.34	2.55	0.0%	0.0%	0.0%	0.0%	26.5%	28.1%	6.8%
10	0.17	-0.25	-0.25	16.1%	1.6%	13.6%	7.4%	10.8%	-2.9%	9.9%
Prediction				1.93	1.68	1.53	1.23	1.98	1.74	1.70
Fit				0.19	0.10	0.41	0.18	0.26	0.14	0.31
Asymmetry				0.00	0.00	0.00	0.02	0.00	0.01	
Adjusted Fit				0.37	0.21	0.41	0.20	0.26	0.14	
Adjusted Fit Weight				23.4%	13.3%	25.3%	12.7%	16.4%	8.9%	

Simulation of Missing Information

Next, we simulate the effect of assigning zero to the relevance weights of observations with missing information within grid prediction. Our simulated dataset comprises a training sample and a testing sample.

- 100 training observations
- 100 testing observations
- 5 normally distributed uncorrelated predictive variables (X) with means equal to 0 and standard deviations equal to 1
- The outcomes (Y) equal the sum of X variables (betas = 1) plus random noise (standard normal)

We consider three training samples.

- Training sample with complete information
- Training sample in which we remove the 25% least informative observations from each predictive variable
- Training sample in which we remove the 25% most informative observations from each predictive variable

We determine informativeness for each variable in isolation as defined by Equation 3.

We calibrate the RBP prediction tasks as follows.

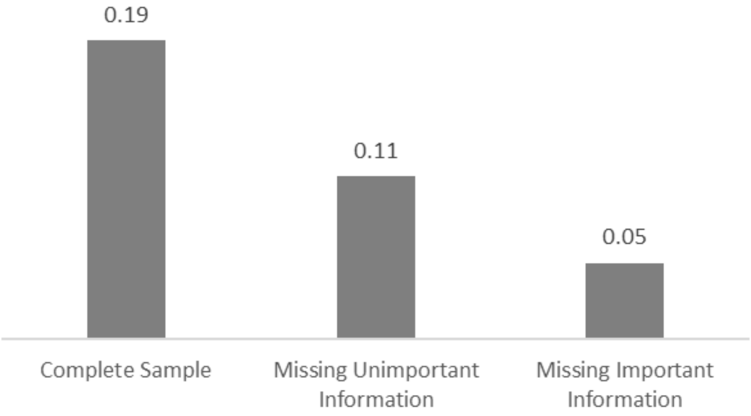
- All 31 combinations of one to five variables
- Relevance thresholds equal to 0.0, 0.2, 0.5, and 0.8
- Censoring criteria: relevance and similarity

The prediction grid for each prediction task, therefore, comprises 248 cells (31 x 4 x 2).

Results

As we should expect from the toy examples presented earlier, predictions have lower average fit when they are missing information, and especially when they are missing important information, as shown in Exhibit 8.

Exhibit 8: Average Fit of Predictions from Complete and Incomplete Training Data



Next, we consider the relationships between the predictions with each other and with the actual outcomes. Exhibit 9 shows that the predictions formed with unimportant missing information are more highly correlated with the actual outcomes than those formed with important missing information and almost as highly correlated with predictions formed from the full sample. Moreover, predictions formed with unimportant missing information are significantly more highly correlated with predictions formed from the full sample than predictions formed with important missing information.

Exhibit 9: Correlations of Predictions with Actual Outcomes and with Each Other

			Missing	Missing
			Unimportant	Important
	Actual	Complete	Information	Information
Actual	1.00			
Complete sample	0.87	1.00		
Missing unimportant information	0.85	0.99	1.00	
Missing important information	0.76	0.90	0.89	1.00

Finally, we analyze how missing information affects the efficacy of predictions for predictions that experienced the 25% greatest decline in fit. The bars on the left of Exhibit 10 show the average difference in absolute errors between predictions and outcomes with unimportant missing information. The bars on the right show the average difference in absolute errors between predictions and outcomes with important missing information. The light bars show the effect across all predictions whereas the dark bars show the effect for those predictions that experienced the 25% greatest decline in fit.

Exhibit 10: Average Difference in Absolute Errors for Predictions with Missing Information

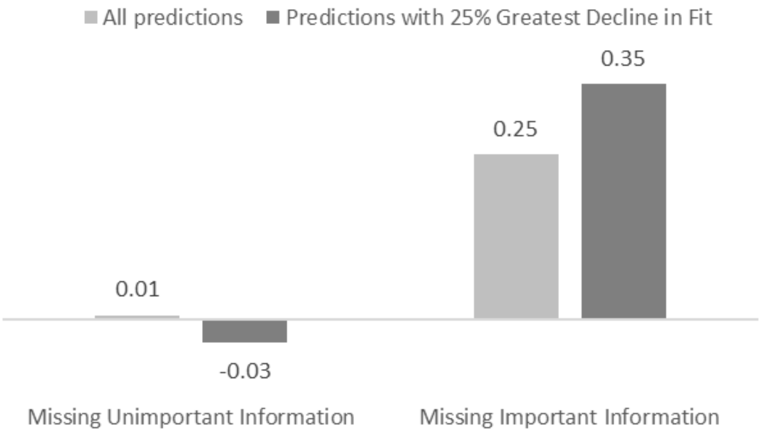


Exhibit 10 reveals several important points.

- On average, predictions formed from observations that are missing important information have larger errors than predictions formed with complete information.

- The prediction errors are significantly greater for predictions that have experienced the greatest declines in fit, which underscores the fact that changes in confidence, which are known in advance, foretell changes in predictive efficacy.
- There are no clear patterns for predictions formed from observations with unimportant missing information.

Summary

RBP is a model-free prediction routine that forms a prediction as a weighted average of observed outcomes in which the weights are based on a precise and theoretically justified statistic called relevance. Unlike a prediction model, which requires an uninterrupted sample of observations to form predictions, RBP seamlessly forms predictions from samples with missing information interspersed throughout the sample by assigning zero to the relevance weights of observations with missing information and blending information across a grid of predictive configurations. RBP's grid prediction evaluates the fundamental tradeoff of missing data whereby predictions formed from more variables tend to have fewer observations, and vice versa.

RBP relies on fit which quantifies the prevalence of useful patterns in a dataset and gives a principled way to anticipate the efficacy of alternative combinations of observations and predictive variables. Fit also gives advance guidance about a prediction's reliability.

Additionally, RBP uses grid prediction to form a composite prediction as a reliability-weighted average of the predictions given by all the combinations of observations and predictive variables.

We showed that the RBP approach for treating missing information preserves more available information than reducing the sample of observations or omitting variables with missing information. We also explained why giving a weight of zero to observations with missing information automatically accounts for the relative importance of the observations. We then presented a toy example to illustrate how the fit of a cell's prediction declines more if the prediction is based on missing information that is important compared to unimportant missing information. We extended our analysis to the composite prediction given by the prediction grid. Again, using toy examples, we showed that the grid's composite prediction is more sensitive to important missing information than unimportant missing information. We concluded by simulating the effect of assigning zero to the relevance weights of observations with missing information. Our simulations offered supportive evidence that RBP's approach for treating missing information reliably accounts for the relative importance of observations with missing information and that it gives advance notice of the effect of missing information on the prediction.

Notes

This material is for informational purposes only. The views expressed in this material are the views of the authors, are provided "as-is" at the time of first publication, are not intended for distribution to any person or entity in any jurisdiction where such distribution or use would be contrary to applicable law and are not an offer or solicitation to buy or sell securities or any product. The views expressed do not necessarily represent the views of Windham Capital Management, State Street Global Markets®, or State Street Corporation® and its affiliates.

References

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022a. "Relevance." *The Journal of Investment Management*, 20 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022b. *Prediction Revisited: The Importance of Observation*. Hoboken, New Jersey: John S. Wiley & Sons.

Czasonis, Megan, Mark Kritzman, and David Turkington. 2023. "Relevance-Based Prediction: A Transparent and Adaptive Alternative to Machine Learning." *The Journal of Financial Data Science*, 5 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2024a. "The Virtue of Transparency: How to Maximize the Utility of Data Without Overfitting." *The Journal of Financial Data Science*, 7 (2).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2024b. "A Transparent Alternative to Neural Networks with an Application to Predicting Volatility." *MIT Working Paper* (September).

Mahalanobis, Prasanta Chandra. 1936. "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Sciences of India*, 2 (1): 49–55.

Shannon, Claude. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal*, 27 (July, October): 379–423, 623–656.

Shapley, L. S. (1953), "A Value for n-Person Games," in *Contributions to the Theory of Games* (Vol. II), eds. H. W. Kuhn and A. W. Tucker, Princeton, NJ: Princeton University Press, pp. 307–318.

[1] Manufacturing replacement information for missing data presents two fundamental problems. First, it requires upfront assumptions that are not informed by the subsequent data analysis, leading to a problem of circularity. For example, it is potentially inconsistent to manufacture replacement information under the assumption of an unconditional relationship where subsequent analysis reveals a strongly conditional relationship, or vice versa. Second, the manufactured data appears as trustworthy as the observed data, which can lead to overconfidence in a prediction.

[2] See, for example, Czasonis, Kritzman, and Turkington (2024a) and Czasonis, Kritzman, and Turkington (2024b).

[3] The descriptions of these concepts follow closely from Czasonis, Kritzman, and Turkington (2022a), Czasonis, Kritzman, and Turkington (2022b), Czasonis, Kritzman, and Turkington (2023), and Czasonis, Kritzman, and Turkington (2024a), and Czasonis, Kritzman, and Turkington (2024b), but they are modified to fit the context of the current discussion.

[4] This measure was first introduced by Mahalanobis (1936).

[5] Shannon showed that information is an inverse logarithmic function of probability, which is a key insight from his comprehensive theory of communication. See Shannon (1948).

[6] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[7] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[8] See Czasonis, Kritzman, and Turkington (2022b) for proof of this result.