The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models

Brian Greenhill Dartmouth College
Michael D. Ward Duke University
Audrey Sacks University of Washington

We present a visual method for assessing the predictive power of models with binary outcomes. This technique allows the analyst to evaluate model fit based upon the models' ability to consistently match high-probability predictions to actual occurrences of the event of interest, and low-probability predictions to nonoccurrences of the event of interest. Unlike existing methods for assessing predictive power for logit and probit models such as Percent Correctly Predicted statistics, Brier scores, and the ROC plot, our "separation plot" has the advantage of producing a visual display that is informative and easy to explain to a general audience, while also remaining insensitive to the often arbitrary probability thresholds that are used to distinguish between predicted events and nonevents. We demonstrate the effectiveness of this technique in building predictive models in a number of different areas of political research.

Binary data are widespread in political science, and political scientists have contributed tremendously to methods for the systematic study of dichotomous variables. Aldrich and Nelson (1984) provided an early, didactic introduction to logit and probit regression. Political science settled on logit, as opposed to probit, largely because of the availability of computer software for the former, and that has become the de facto standard. However, until recently, regression results from these binary, discrete regression models were mainly presented as tables of coefficients and associated measures of precision. Because these numerical results are so difficult to interpret, and so prone to misinterpretation, King, Tomz, and Wittenberg (2000) introduced a more graphical way of presenting the results in terms of calculating expected values conditional on the estimated model. The basic ideas presented therein serve as the basis for the generic approach to presenting results found in the omnibus software package known as *Zelig* (Imai, King, and Lau 2008).

Despite these improvements, very scant attention has been paid to the notion of the fit of models from a more modern and visual perspective. Until recently, most inferential problems in political science were entirely theoretical and retrospective: looking at statistical significance and looking at observed data. However, today there is considerably greater interest in cross-validation and out-of-sample prediction than ever before. In addition, there is a growing movement that strives to present results in a visual, intuitive fashion, rather than an analytical and

Brian Greenhill is an Assistant Professor of Government at Dartmouth College, 223 Silsby Hall, HB 6108, Hanover, NH 03755 (brian.d.greenhill@dartmouth.edu). Michael D. Ward is a Professor of Political Science at Duke University, 408 Perkins Library, Campus Box 90204, Durham, NC 27708 (mw160@duke.edu). Audrey Sacks is a Ph.D. candidate in the Department of Sociology, University of Washington, 211 Savery Hall, Box 353340, Seattle, WA 98195-3340 (sacks@u.washington.edu).

An earlier version of this article was presented at the 2009 annual meeting of the American Political Science Association, Toronto, Canada. We appreciate the comments and suggestions we received from John Ahlquist, Hyeran Jo, Francisco Pedraza, and from colleagues associated with the ICEWS project, especially Sean O'Brien, Philippe Loustanau, and Laura Stuart. Colleagues at the 10th Anniversary Conference on Statistics and the Social Sciences, at the University of Washington, Seattle, WA, also provided useful input, especially Andrew Gelman and Steve Fienberg. We also presented a version of these ideas at the 2009 Summer Political Methodology Conference at Yale University and received valuable reactions from a variety of colleagues. Despite all this good advice, we remain responsible for results presented herein. This project was partially supported by the Information Processing Technology Office of the Defense Advanced Research Projects Agency aimed at producing models to provide an Integrated Crisis Early Warning Systems (ICEWS) for decision makers in the U.S. defense community. The holding grant is to the Lockheed Martin Corporation, Contract FA8650-07-C-7749. The anonymous reviewers and editor made exceptionally helpful comments. The views presented in this article are those of the authors, not those of any sponsors. Replication data for this article are available at http://dvn.iq.harvard.edu/dvn/dv/briangreenhill.

American Journal of Political Science, Vol. 55, No. 4, October 2011, Pp. 990–1002

DOI: 10.1111/j.1540-5907.2011.00525.x

numerical one. Much of this was stimulated by Tufte's work (2006), but the cause has deep roots in statistics as well (Gelman, Pasarica, and Dodhia 2002), and more recently has been promoted in political science (Kastellec and Leoni 2007). As a result, the die-hards from the classical perspective who wanted "some measure of fit" have been joined by a wide variety of other interests who desire more information about the quality of empirical models than is provided by tables of numbers representing the means and variances of the estimated parameters. There is now greater interest in ROC plots and tables of specificity and sensitivity as a way of gauging the validity of the estimated model, in empirical terms.

Predicting Dichotomous Outcomes

Models with dichotomous outcome variables present a particular challenge for the assessment of model fit. This is because these models generate fitted values that lie somewhere along a continuous $0 - 1$ scale (e.g., $\hat{p} = 0.68$), whereas the actual values of the dependent variable are dichotomous (for all observations, $y = 1$ or $y = 0$). This raises the question of what rule should be applied for comparing the probabilities to the actual outcomes. To put it in more concrete terms, if we have a logit model of U.S. presidential election outcomes that generates predicted probabilities of 0.68 and 0.32 for Obama and McCain victories in the 2008 presidential election, how "correct" is the model? A common choice is to simply say that the model makes a correct prediction in this instance because the model estimates a probability of an Obama win that is greater than 0.5. But applying a threshold like this has the effect of collapsing the differences in model fit that exist between a relatively poor model that assigns a probability of only 0.51 to an Obama victory and a superior model that assigns a probability of 0.99. Before presenting our new visual method—what we call the "separation plot"—for dealing with this problem, we shall briefly review the other heuristics that are commonly used to assess the predictive power of logit or probit models.

We begin by demonstrating how these alternative heuristics are used for the small ($N = 6$) dummy data set that we provide in Table 1. We know that no one will actually undertake this kind of analysis with only six data points, but it provides a simple pedagogical framework for explaining the intuition of our approach. Let's suppose that the data represent instances of civil war within a sample of six countries. Our dependent variable, y, is coded such that each instance of war is assigned a value of 1 and each instance of peace a value of 0. The fitted values, $\hat{p}$, obtained from a logit or probit model are shown in the third column. As the table shows, in some cases the model assigns a high probability to cases of actual war—e.g., Country C with a $\hat{p}$ of 0.997—and therefore does a relatively good job of fitting the data. In other cases—e.g., Country F with a $\hat{p}$ of 0.422—the model seems to perform less well. The challenge, therefore, is to find a simple yet informative way of summarizing the model's overall degree of fit.

TABLE 1 Sample Data

Country	Actual Outcome (y)	Fitted Value ($\hat{p}$)
A	0	0.774
B	0	0.364
C	1	0.997
D	0	0.728
E	1	0.961
F	1	0.422

Contingency Tables

One commonly used heuristic is the contingency table. This is a table created by dichotomizing the predicted probabilities. As mentioned in the electoral example above, we could establish a rule whereby every fitted value less than 0.5 is deemed an instance of peace, whereas every value greater than or equal to 0.5 is analogously considered an instance of war. Using the data on six hypothetical countries given in Table 1, we can generate the following contingency table:

	Predicted War	Predicted Peace
Actual War	{C, E}	{F}
Actual Peace	{A, D}	{B}

According to this scheme, we find that three countries were correctly predicted—C, E (correct predictions of War), and B (Peace)—whereas the three others were incorrectly predicted—A, D, and F. This allows us to calculate a "Percentage Correctly Predicted" (PCP) of 50%. But these results, of course, depend upon our choice of threshold. After all, 0.5 is an entirely arbitrary threshold and were one to choose a threshold of, say, 0.3, the results would look rather different:[1]

[1] In this example, it happens that the PCP remains at 50% when the threshold is lowered to 0.3; however, this is not the case for other thresholds. For example, a threshold of 0.4 causes the PCP to rise to 67%.

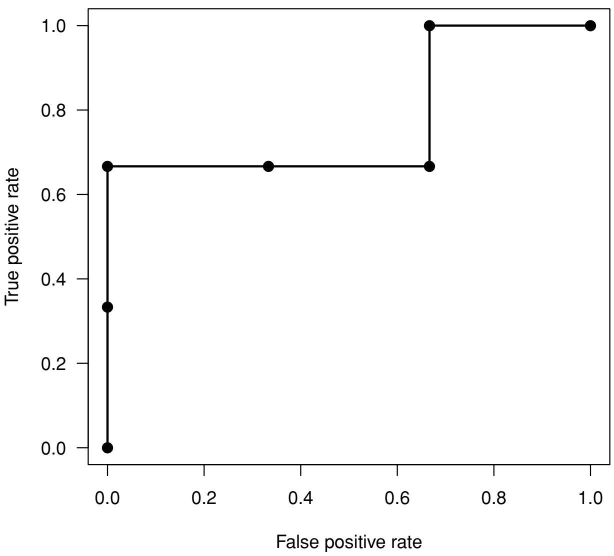
	Predicted War	Predicted Peace
Actual War	{C, E, F}	{}
Actual Peace	{A, B, D}	{}

By lowering the threshold to 0.3, we are now able to predict war in three of the countries where there actually was a war—C, E, and F—but now we don't predict any instances of peace. Instead, we incorrectly predict war in the three actual cases of peace—A, B, and D. By lowering the threshold from 0.5 to 0.3, we have increased our number of true positives at the cost of a higher number of false positives. While these 2 × 2 tables are easy for readers to understand, this simplicity comes at the cost of (1) the loss of much of the information contained in the model's output (e.g., no distinction is made between probabilities of 0.51 and 0.99 when the threshold is set at 0.5); and (2) the author having to choose a particular threshold often without having any theoretical justification for doing so. There may be some applications where the threshold is not arbitrary. For example, if we were to decide that when it comes to predicting wars, a false negative is 10 times more costly than a false positive, then a threshold of 0.1 might be deemed appropriate. But generally, in political science we do not have such strong theoretical guidance about the cost function that separates false positives from false negatives.

The ROC Curve

By now a widely used heuristic that avoids the problem of arbitrary thresholds is the Receiver Operating Characteristic (ROC) curve. The so-called ROC curve provides a visual summary of the trade-off between false positives and false negatives as the threshold is varied. Usually it is presented in the form of a plot of the false positive rate against the true positive rate obtained for each possible threshold, τ. The false positive rate (FPR) is defined as the number of false positives divided by the sum of the false positives and true negatives (in other words, all of the incorrectly identified negatives divided by all of the actual negatives), whereas the true positive rate (TPR) is defined as the number of true positives divided by the sum of the true positives and false negatives (all of the correctly identified positives divided by all of the actual positives). The True Positive Rate is also referred to as the *sensitivity* of a classifier. The False Positive Rate is equivalent to 1—*specificity*, where specificity is defined as $\frac{TN}{FP+TN}$. Thus, the ROC curve is essentially a map of the sensitivity versus the specificity for different cost functions.

FIGURE 1 ROC Plot for the Data Shown in Table 1



The ROC curve for our six-row data set on war and peace is shown in Figure 1. In this case, the six discrete $\hat{p}$ values give rise to seven possible combinations of FPR/TPR values, and therefore seven points on the curve. The calculation of all seven points on the curve is shown in Table 2.

ROC curves have the advantage of providing a visual description of the predictive power of the model over all possible thresholds. Models with high levels of predictive power will tend to have true positive rates that are consistently higher than the corresponding false positive rates, giving rise to curves that have the appearance of being pulled toward the upper-left corner of the plot. As a result, the overall predictive power of the model (across all possible thresholds) can be summarized in terms of the area under the ROC curve since the ROC is defined on the unit square. The area under this curve is denoted the "AUC score." AUC scores are bounded between 0 and 1. A model that is no better than a simple coin toss should have an AUC of 0.5; lower scores are occasionally observed when, on average, the false positive rate *exceeds* the true positive rate, indicating a model that produces predictions worse than you could expect by chance. However, the value of ROC curves is limited by the fact that the particular shape of the curve tells us little about the model's fit, and in practice it is only the total area enclosed by the curve—i.e., the single number statistic provided by the AUC score—that allows us to get a sense of the model's fit.

TABLE 2 Calculation of the Points on the ROC Curve

Threshold	TP	FP	FN	TN	FPR	TPR
$0 < \tau < 0.364$	{C, E, F}	{A, B, D}	{}	{}	$\frac{3}{3+0} = 1$	$\frac{3}{3+0} = 1$
$0.364 < \tau < 0.422$	{C, E, F}	{A, D}	{}	{B}	$\frac{2}{2+1} = 0.67$	$\frac{3}{3+0} = 1$
$0.422 < \tau < 0.728$	{C, E}	{A, D}	{F}	{B}	$\frac{2}{2+1} = 0.67$	$\frac{2}{2+1} = 0.67$
$0.728 < \tau < 0.774$	{C, E}	{A}	{F}	{B, D}	$\frac{1}{1+2} = 0.33$	$\frac{2}{2+1} = 0.67$
$0.774 < \tau < 0.961$	{C, E}	{}	{F}	{A, B, D}	$\frac{0}{0+3} = 0$	$\frac{2}{2+1} = 0.67$
$0.761 < \tau < 0.997$	{C}	{}	{E, F}	{A, B, D}	$\frac{0}{0+3} = 0$	$\frac{1}{1+2} = 0.33$
$0.997 < \tau < 1$	{}	{}	{C, E, F}	{A, B, D}	$\frac{0}{0+3} = 0$	$\frac{0}{0+3} = 0$

TABLE 3 Calculation of Brier Scores

Country	Actual Outcome (y)	Fitted Value ($\hat{p}$)	Brier Score $(\hat{p} - y)^2$
A	0	0.774	0.599
B	0	0.364	0.132
C	1	0.997	0.000
D	0	0.728	0.530
E	1	0.961	0.002
F	1	0.422	0.334

Brier Scores

Another common choice of summary statistic is the Brier Score, which is the mean value of the squared difference between the fitted and actual values of the dependent variable. The Brier score was developed in the early 1950s to provide a way to grade probabilistic weather forecasts (Brier 1950). The formula for this metric is quite simple:

$$B = \frac{1}{N} \sum (\hat{p} - X)^2$$

where p is the probabilistic forecast, X is the dichotomous, binary variable of whether or not the forecast event occurred (1 = yes; 0 = no), and N is the number of observations. The closer to zero the Brier score, the better the forecasts. As is shown in the calculations in Table 3, in the case of our civil war example the Brier score is 0.266.

Expected PCP

Herron (1999) suggests an alternative to the traditional calculation of Percentage of Correct Predictions (PCP) in a way that avoids using an arbitrary threshold to convert the predicted probabilities ($\hat{p}$) into fitted values of the dependent variable (y). The "expected PCP" (ePCP) is calculated as follows (where N is the total number of observations):

$$ePCP = \frac{1}{N} \left(\sum_{y_i=1} \hat{p}_i + \sum_{y_i=0} (1 - \hat{p}_i) \right).$$

The ePCP is essentially a measure of the average of the probabilities that the model assigns to the correct outcome category for each observation (whether that may be 0 or 1). It has the advantage of being easily extended to measuring the fit of categorical models with more than two outcome categories. The ePCP for our illustrative example can be calculated as follows:

$$\begin{aligned} ePCP = \frac{1}{6} & (0.997 + 0.961 + 0.422 + (1 - 0.774) \\ & + (1 - 0.364) + (1 - 0.728)) = 0.586. \end{aligned}$$

This statistic tells us that, on average, our hypothetical model therefore assigns 59% of the probability density to the correct outcome category. This metric provides a heuristic that actually has a statistical interpretation, something that is not the case for most other single number summaries.

Pseudo R^2

Sometimes scholars (and software) will report a *Pseudo* R^2 which is typically one minus the ratio of the likelihood for a null model to the likelihood for the estimated model, so that if the null model and the estimated model have about the same likelihood, then the pseudo R^2 score is close to zero. Like the R^2 this measure has many flaws as a measure of fit, but it does provide a widely used single number summary of the fit of discrete models (McKelvey and Zavoina 1975).

The Separation Plot

To help provide a more nuanced and nonscalar, visual yardstick for the performance of such models, we present

TABLE 4 Rearrangement (and Coloring) of the Data Presented in Table 1 for Use in the Separation Plot

Country	Fitted Value ($\hat{p}$)	Actual Outcome (y)
B	0.364	0
F	0.422	1
D	0.728	0
A	0.774	0
E	0.961	1
C	0.997	1

a new approach that we call the separation plot. We shall first explain how the plot is produced and then consider some of the advantages it has over alternative methods of assessing model fit. In the next section, we show how the separation plot can be used to evaluate the fit of a number of widely cited models in political science.

The Concept

Construction of the separation plot begins by simply rearranging the data presented in Table 1 such that the fitted values are presented in ascending order. We then note whether each of these corresponds to an actual instance of the event (war) or a nonevent (peace). The results of this rearrangement are shown in Table 4. Here we have highlighted the rows corresponding to actual events in red (or gray if viewed in black-and-white), and the rows corresponding to nonevents in light yellow (or white).

The key idea is that the model's fit (or predictive power) can now be evaluated by simply gauging the extent to which the actual instances of the event are concentrated at the bottom end of the table, and the nonevents at the top end of the table. A model with no predictive power—i.e., one whose outcomes can be approximated by a random coin toss—would generate an even distribution of 0s and 1s along the column on the right-hand side. On the other hand, a model with perfect predictive power would produce a complete separation of the 0s and 1s in the right-hand column: low fitted values would always turn out to be associated with actual instances of peace (0s), whereas high fitted values would always be associated with actual instances of war (1s). A perfect model would therefore show complete separation of the dark and light-colored rows in Table 4.[2]

FIGURE 2 Separation Plot Representing the Data Presented in Table 1

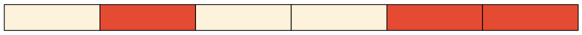
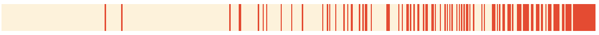


FIGURE 3 Separation Plot for a Larger Data Set



It turns out to be very easy to discern these differences using the simple graphical representation that we call the "separation plot" (see Figure 2). In this graph, the dark and light panels correspond to the actual instances of the events and nonevents, respectively,[3] ordered such that the corresponding $\hat{p}$ values increase from left to right.[4] (One can think of this graph as simply a 90° rotation of Table 4.) As Figure 2 shows, our "model" does a reasonably good job of describing the data given that most of the events are clustered on the right-hand side of the graph. Remember that a perfect model would produce a plot where all the events are clustered at the right-hand pole and all the nonevents at the left-hand pole, i.e., ⬜⬜⬜🟥🟥🟥, whereas a completely ineffective model that shows no such separation taking place might look more like 🟥⬜🟥🟥⬜⬜.[5]

Adding Information

Figure 3 shows a separation plot obtained from fitting a simple bivariate logit model to a larger data set that consists of 500 observations. (In this case, a model has

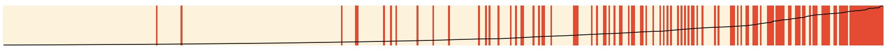
[2] We would like to thank one of the *AJPS* anonymous reviewers for pointing out the similarity that exists between the separation plot and a technique used in molecular genetics called "Gene Set Enrichment Analysis" (GSEA). In GSEA, the ranked differences in expression levels of a large number of genes associated with two different phenotypes are matched to a set of genes that are known to be associated with a specific biological process. If this gene set appears to cluster among the genes that show the highest level of difference between expression in the two phenotypes—i.e., at the upper end of the plot—it is more likely that this particular gene set is associated with the difference in phenotype. For more details, see Subramanian et al. (2005). DNA bar coding uses a similar approach.

[3] The colors used for this plot were obtained from one of the color schemes developed by Harrower and Brewer (2003).

[4] A question that sometimes arises is how one should deal with ties in the levels of $\hat{p}$. The R function that we have written to generate separation plots gives the user the option of dealing with ties by either (1) including these observations in the same order in which they appeared in the original data set or (2) randomizing their order in an attempt to reduce the possibility of seeing a pattern in the separation plot that is simply an artifact of the ordering of observations in the original data set.

[5] Worse still, a model that consistently makes the *wrong* predictions would produce a plot looking like 🟥🟥🟥⬜⬜⬜ —i.e., one in which the events and nonevents have been sorted in the wrong direction.

Figure 4 **Adding a Graph of $\hat{p}$ to the Separation Plot**



been fitted to a simulated data set in which a correlation of 0.55 was induced between the response and explanatory variables.) Here we can see that the degree of separation is high, suggesting that the model makes a reasonably good fit to the data. It turns out that the area under the ROC curve is 0.87.

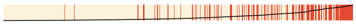
We can now add more information to the plot by including a line that represents the values of $\hat{p}$ that we had estimated for each of the observations using our model (see Figure 4). In this case, we can imagine the existence of a y-axis where the lower edge of the plot represents $\hat{p} = 0$ and the upper edge represents $\hat{p} = 1$. Having reordered the cases by their values of $\hat{p}$ in the first step, the line will of course slope upwards from left to right. The inclusion of this line allows us to examine the correspondence that exists between the levels of $\hat{p}$ and the levels of the dependent variable. In doing so it allows us to see whether the overall degree of separation between events and nonevents is associated with sharp differences in the level of $\hat{p}$, or more modest differences. Figure 4 shows that very few observations where $y = 1$ (i.e., the darker lines) are found among the low values of $\hat{p}$ (i.e., the left-hand side of the graph), whereas a much greater concentration of actual events is found among the very high values of $\hat{p}$.

Another quantity of interest is the expected number of total events predicted by the model. We can calculate this by simply adding up the predicted probabilities across all observations. (This value can also be thought of as the integral of the $\hat{p}$ line added to the separation plot in Figure 4.)

$$\textit{Expected Number of Events} = \sum_{i=1}^{N} \hat{p}_i$$

The expected number of events for the simulated data in this section is 118. We can represent this quantity on the separation plot by adding a marker at the 118th highest value of $\hat{y}$. This is shown by the small triangle added to the bottom of Figure 5. This allows the user to see how the total number of events predicted by the model compares to the actual number of events in the data. As is illustrated by Figure 6, a model that fits the data perfectly will not only show full separation between events and nonevents, but will also predict a total number of events that is equal to the total number of actual events that occur in the data.

Advantages

The main advantage of the separation plot is that it provides a quick and easy-to-understand summary of the fit of a model without the loss of information associated with a single number statistic. Moreover, it can be easily and compactly incorporated into a section of text as an in-line graphic (what Tufte 2006 refers to as a “sparkline”) like this: . Embedded within the separation plot are the following pieces of information:

1. The relative number of 0s and 1s in the actual data. This provides an indication of the sparsity of the events in the data used to generate the model;
2. The range and degree of variation among the predicted probabilities generated by the model;
3. The degree to which high predicted probabilities correspond to actual instances of the event, and low predicted probabilities correspond to nonevents—in other words, a visual summary of the fit of the model;
4. The total number of events predicted by the model.

The second point is especially important given that some logit or probit models assign probabilities to outcomes that fall within a surprisingly narrow range. For example, the maximum probability assigned by the main model (“Model 1”) of civil war onset in Fearon and Laitin (2003) is only 0.48, and only two of the 6,237 observations used to estimate that model are assigned a probability of greater than 0.3. As is demonstrated in the following section, the separation plot allows us to see directly how these predicted probabilities map on to the actual outcomes of the event of interest. (As it turns out, only one of these two predicted probabilities above 0.3 corresponds to an actual instance of civil war onset.)

The separation plot can be especially helpful at the model selection stage, where single-number summaries of predictive power may lack nuance and ROC curves are more difficult to compare. Consider the following example of an attempt to build a model that describes the

FIGURE 5 Adding the Expected Number of Events

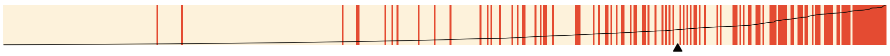


FIGURE 6 A "Perfect" Model for the Same Data Used in Figure 5

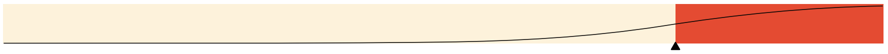
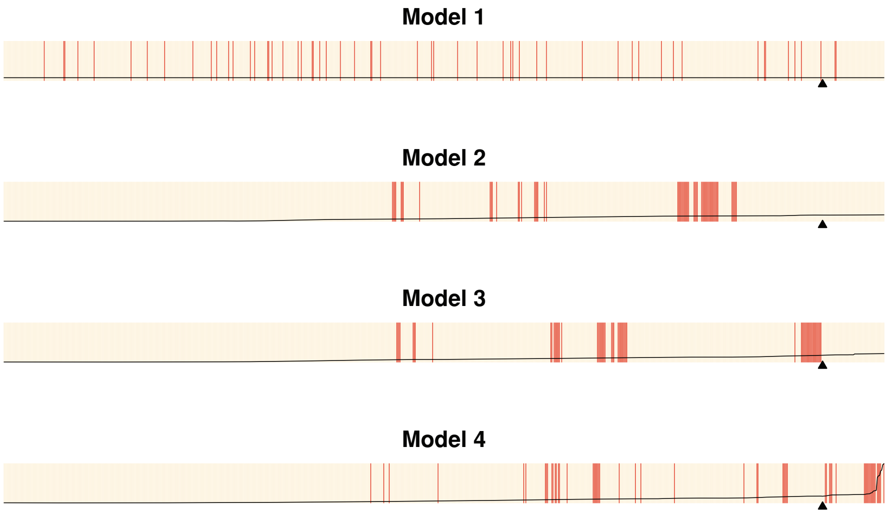


FIGURE 7 Separation Plots Used in the Development of a Model of Insurgency in the Asia-Pacific Region, 1998–2004



Note: For comparison, Models 1–4 have AUC scores of 0.500, 0.714, 0.744, and 0.816; Brier scores of 0.065, 0.063, 0.062, and 0.057; and ePCP scores of 0.869, 0.875, 0.876, and 0.887.

incidence of political insurgencies among a group of 29 countries in the Asia-Pacific region over the 1998–2004 period ($n = 812$). The project and data are described in more detail in O'Brien (2010). Figure 7 shows the results of using separation plots at successive stages of the model selection process. In the first graph, we show the results of fitting a model that has an intercept and no covariates. In this case, the fitted values are identical for every observation in the data set and are simply the mean of the insurgency variable. The model clearly does a poor job of separating events from nonevents, and the line representing the corresponding values of $\hat{p}$ remains constant across all observations.

In the second graph we add a single covariate, GDP per capita (lagged by one year). This clearly improves the model's ability to distinguish events from nonevents. We now see three distinct clusters of events arranged toward the right-hand side of the plot space (corresponding to the higher values of $\hat{p}$). Meanwhile, no events are detected in the regions corresponding to the lower ∼40% of observations.

The third graph shows the results of adding a second covariate, anocracy, which can be thought of as an indicator of the absence of any effective government (Marshall and Jaggers 2003). Adding this variable clearly improves the predictive power of the model: the clusters of dark lines indicating actual instances of the event are shifted further to the right, and no events appear to the left of the leftmost insurgency event in the second graph.

Finally, the graph at the bottom of Figure 7 shows the effect of including two further variables representing detailed event stream data on the number of instances of hostility between the government and the insurgent groups in the preceding period. In this case, we see a distinctive clustering of observations corresponding to insurgency events on the far right-hand side of the graph, and we can also see that these are associated with significantly higher values of $\hat{p}$.

These results show how the separation plots can be used to gain a more fine-grained understanding of the way in which the in-sample fit or out-of-sample predictive power of the model improves as one changes the model specification. One of its major advantages over a single statistic such as area under the ROC curve is that it allows the user to consider gains and losses in predictive power within different regions of the data. For example, it may be the case that the user is especially interested in developing a model that maximizes the number of correctly predicted events among the highest values of $\hat{p}$, in which case he or she could focus only on the region on the far right-hand side while ignoring other parts of the graph. On the other hand, certain applications may call for the user to focus on minimizing the number of events among lower values of $\hat{p}$, in which case the third model in Figure 7 would actually be considered superior to the fourth.[6]

Another important advantage of the separation plot in the model-fitting stage is that it allows for the identification of clusters of false negatives and/or false positives in the data. These would be indicated by a cluster of events at low values of $\hat{p}$ or a cluster of nonevents at high values of $\hat{p}$, respectively. Moreover, this has the additional benefit of bringing our attention to unusual events or possible coding errors in the data.[7]

Examples

We turn now to the examination of a number of examples of models of dichotomous outcomes that have been used in different subfields of political science. In each case, we show how the separation plot can be used to enhance our understanding of the model's fit.

Political Campaigns: Hillygus and Jackman (2003)

Hillygus and Jackman (2003) develop a model of voting intentions that shows not only that campaign events such as the party conventions and presidential debates lead to changes in voter preferences, but also that the effect that these events have on voter preferences varies according to the preferences that the voters expressed in an earlier survey. They interpret this finding as evidence that voters assimilate new information about the candidates in a way that is conditional upon their earlier preferences (Hillygus and Jackman 2003, 590).

The authors employ a logit model of voter preference. In addition to discussing the estimates of the coefficients in the model, they also comment on the fit of the model in the following way:

> In gross terms, the estimated transition models fit the data very well, as does any model with a lagged dependent variable in the equation. The convention model correctly predicts 91% of vote preferences, and the debate model correctly predicts 94% of vote preferences, using $p = .5$ as the classification threshold. Moreover, each of the models has an area under the ROC curves anywhere from .79 to .89. These numbers indicate that each of our models does an acceptable to excellent job of discriminating those who transitioned from those who were stable. (Hillygus and Jackman 2003, 592–93)

We believe that the separation plot provides a more informative yet compact tool for conveying the same point. As is shown in the separation plots in Figure 8, the models appear to make an excellent fit to the data. Moreover, unlike the summary statistics presented in the original article, the separation plots show that the total number of events in the data—i.e., statements of intentions to vote for a particular candidate—is very large.

Civil War: Fearon and Laitin (2003)

Fearon and Laitin (2003) developed a highly influential model of the probability of civil war onset in the post-WWII period. Based upon the statistical significance of the variables included in the logistic regression model, the authors claim that various economic and geographical factors that favor insurgencies (e.g., poverty, large populations, and mountainous terrain) tend to be associated with the onset of civil war. Most importantly, they find that cultural factors such as the extent of ethnic heterogeneity are not associated with higher probabilities of civil war onset.

However, when the fit of their main model ("Model 1" in their article) is visualized using a separation plot, the model appears to make a relatively poor fit to the data. Indeed, as we show in Figure 9, this model fits the data only marginally better than an extremely parsimonious

[6] In the R package for the separation plot, we also provide a facility to highlight individual observations of interest. This allows the user to see how the predicted probability assigned to one or more critical cases changes under different specifications of the model.

[7] We are especially grateful to one of the anonymous *AJPS* reviewers for making this observation.

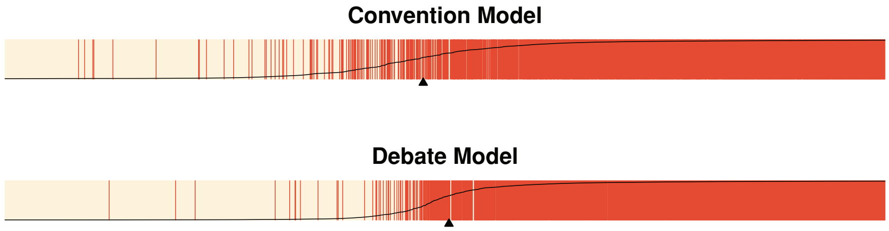
model that includes GDP per capita (logged) as its only covariate. This is a point that is developed in greater detail in Ward, Greenhill, and Bakke (2010).

What Can Go Wrong?

In cases where the sample size is very large ($N \approx >5{,}000$), it can sometimes be difficult to distinguish between the events and nonevents in the separation plot. This occurs because the resolution of a typical screen or printed display is not high enough to allow the user to discern the individual lines representing events and nonevents and can be especially problematic when the events occur rarely in the data. For example, the Fearon and Laitin (2003) model of civil war discussed earlier consists of 6,327 observations, of which only 106 involve cases of civil war. When plotted as a separation plot in the usual way, the events and nonevents have a tendency to blend into one another, as shown by the separation plot in the upper panel of Figure 10. In this case, the width of the individual lines is smaller than the resolution of most printers, and the relatively small number of events in the data (in this case, the 106 instances of civil war) become buried under the the lighter-colored lines representing nonevents.

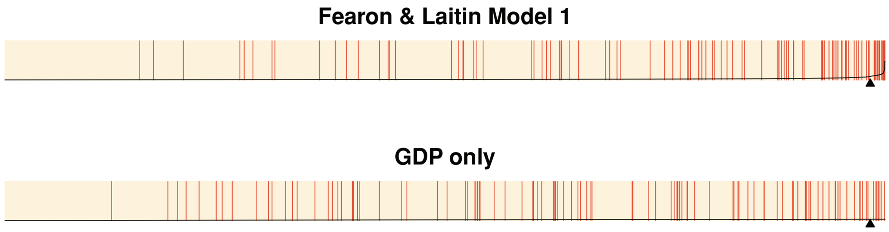
One solution is to use thicker lines for each observation and to add the lines representing the events after all of the lines representing the nonevents have already been plotted. This is what we had done in Figure 9 earlier, which has been reproduced in the middle panel of Figure 10. This has the effect of increasing the prominence of the lines representing the events but necessarily causes some of the nearby nonevents to be obscured. We would therefore recommend interpreting a separation plot like this in conjunction with one produced that increases the thickness of the nonevents relative to the events. This allows the user to see the extent to which nonevents are appearing among clusters of events (especially on the right-hand side of the separation plot—see the lower panel of

FIGURE 8 Separation Plots for the Hillygus and Jackman (2003) Models of Voting Intentions in the 2000 Presidential Election



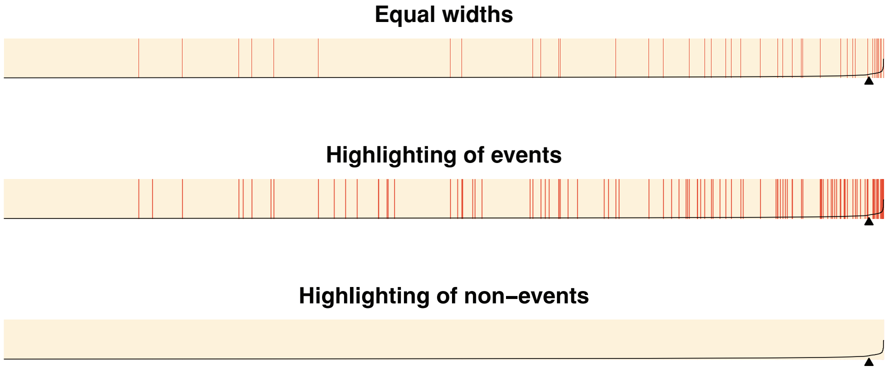
Note: The upper plot shows the results of the survey conducted in the period following the party conventions, while the lower plot shows the results of the survey conducted after the presidential debates. Both models make an excellent fit to the data. (For comparison, the convention and debate models have AUC scores of 0.964 and 0.982; Brier scores of 0.071 and 0.045; and ePCP scores of 0.859 and 0.909.)

FIGURE 9 Comparison of the Separation Plots Produced by Replicating Model 1 of Fearon and Laitin (2003) and by Reestimating the Model with Logged GDP per Capita as the Only Covariate



Note: For comparison, Model 1 and the GDP-only model have AUC scores of 0.760 and 0.671; Brier scores of 0.016 and 0.016; and ePCP scores of 0.968 and 0.967.

FIGURE 10 **Problems (and Potential Solutions) for Separation Plots When the Number of Observations Is Very High Relative to the Size of the Plot**



Note: The upper panel shows a separation plot of the Fearon and Laitin model where the event and nonevent lines are of equal width. The second panel shows the same data plotted in a way that emphasizes the events relative to the nonevents. The third panel emphasizes the nonevents relative to the events.

Figure 10). In this case, the pair of plots allows us to see that while a large proportion of the instances of civil war are clustered on the right-hand side of the plot (where values of $\hat{p}$ are higher), this cluster also contains a large number of nonevents. The R function that we have written to generate separation plots allows the user to choose between these alternative types of display (see appendix).

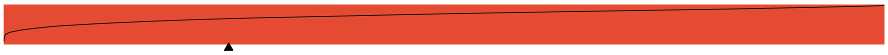
There are a variety of great sources about the ins and outs of displaying and printing information, down to the level of the pixel. In particular, a good overview is found in Tufte (2006, esp. 60–63). Even more detailed information may be found in Sedgewick's work on algorithms (Sedgewick 1998; Sedgewick and Wayne 2011). There are at least two important issues. One is the aspect ratio, and here we have followed Tufte's rule that the length must be greater than the height, but at the same time we have allowed users to specify a ratio of their own. Printing and displaying these graphics is complicated, owing to the different density of the final delivery device (monitor or page), as well as the methods of rendering. Color printers don't necessarily use the same methodology for printing that book publishers use, for example. Images on a 26in LED display will be rendered differently than on smaller LCDs, as well. We have worked to the best of our ability to ensure that the printed and displayed images from the package will work for a wide range of applications. But it should be pointed out that these images are meant to be presented in a very high resolution medium, such as that found in good typography. To date, the examples that have been printed in published work have been relatively artifact absent (Ahlquist 2010). Finally, we have experimented with black-and-white versions of the separation plot, ones that can be printed without extensive page costs. Following the recommendations of Tufte (2006, 62), we suggest a grayscale color scheme that reduces the "optical noise" that tends to arise when closely spaced black-and-white lines are plotted close to each other. This option is implemented in our R package.

Example: Voting Behavior (Rosenstone and Hansen 1993)

In some cases where N is much larger and the degree of separation is very low (i.e., events and nonevents are evenly distributed along the length of the plot), it may still be difficult to detect any meaningful degree of separation using the above techniques. In cases such as these, we suggest using an alternative version of the separation plot that separates the events and nonevents into two separate plots. The objective is again to visually determine whether the events are associated with higher predicted probabilities than the nonevents. We do this by grouping the predicted probabilities into bins and comparing the number of events and nonevents contained within each bin.

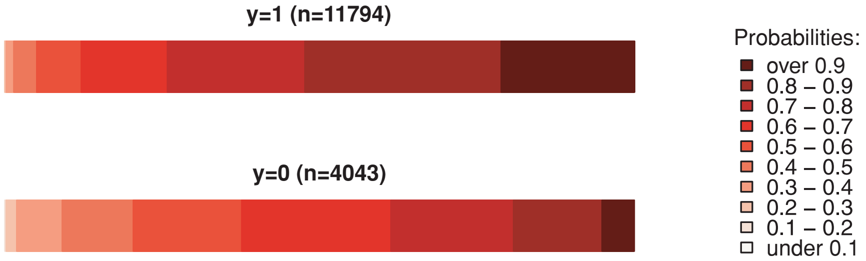
As an example, we draw on Rosenstone and Hansen (1993) and King, Tomz, and Wittenberg (2000). Rosenstone and Hansen (1993) employ a logit model to explain why some individuals are more likely than others to vote in U.S. presidential elections. King, Tomz, and Wittenberg (2000) then use this model as a means to illustrate how to improve the statistical interpretation of regression analyses. For expository purposes, King, Tomz, and

FIGURE 11 **Separation Plots for the Rosenstone and Hansen (1993) and King, Tomz, and Wittenberg (2000) Model of Voter Turnout in Presidential Elections between the Years 1960 and 1996**



Note: The individual lines are impossible to distinguish at this scale when *N* is so large.

FIGURE 12 **An Alternative Version of the Separation Plots for the Rosenstone and Hansen (1993) and King, Tomz, and Wittenberg (2000) Model of Voter Turnout in Presidential Elections between the Years 1960 and 1996**



Note: The upper plot shows the results for cases of actual turnout, while the lower plot shows the results for cases of absenteeism.

Wittenberg (2000) focus only on the following demographic variables that Rosenstone and Hansen (1993) emphasize for explaining voter turnout in presidential elections from the years 1960 to 1996: education, income, age, and race (whites and nonwhites). We replicate the model of voter turnout described above and use the predicted and actual values to create a separation plot.

As illustrated in Figure 11, when we display the results of this model as a standard separation plot the individual dark and light lines cannot be distinguished at this scale because of the large number of observations ($N = 15{,}837$) of this study. We therefore present the alternative version of the separation plot in Figure 12 that groups probabilities into discrete bands as described above. In this graph, the colors correspond to ranges of probabilities; the darkest shade corresponds to a probability of 0.9 and higher and the lightest shade corresponds to a probability of 0.1 and lower.

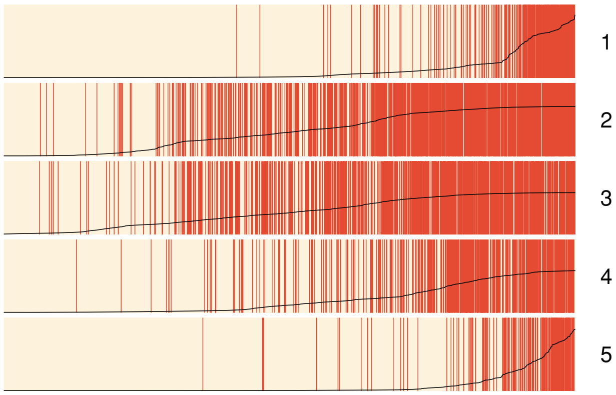
This alternative version of the separation plot suggests that the model performs reasonably well in assigning high probabilities to actual cases of voter turnout. The darker bands representing the high probabilities of the event occurring are considerably wider on the upper deck of the plot (which consists of the actual events) than in the lower deck (which consists of the nonevents). Likewise, the lighter bands corresponding to low probabilities of the event are wider in the lower deck than they are in the upper deck.

Polytomous Dependent Variables

The concept of the separation plot can be extended to assist in the analysis of ordered probit/logit or multinomial models that have more than two categorical outcomes. In these cases, an individual separation plot is generated for each category of outcome. Each plot therefore compares the probabilities of obtaining one particular outcome against a dichotomous indicator of whether or not each case is in fact associated with that particular category of outcome. The user can then review an array of separation plots that represent the predictive power of the model across all possible categories of outcome.

In Figure 13 we show an example of how separation plots can be used to assess the predictive power of an ordered probit model. In this figure we replicate the results of Neumayer (2005, Table 2, Model 6), which examines the effect that ratification of the International Covenant on Civil and Political Rights has on states' human rights practices. The dependent variable is the 5-point

FIGURE 13 **Separation Plots for Each of the Five Levels of Personal Integrity Rights Estimated in Neumayer (2005, Table 2, Model 6)**



Political Terror Scale, which measures the extent to which governments engage in human rights violations aimed at disrupting nonviolent political activities (Wood and Gibney 2010). Higher scores on the Political Terror Scale represent greater levels of repression.

As the graphs show, on the whole Neumayer's model does a good job of matching high probabilities of each outcome to actual occurrences of each outcome. For example, the first separation plot in Figure 13 shows that most of the highest predicted probabilities of generating an outcome of 1 on the Political Terror Scale do in fact correspond to actual outcomes of 1. For the intermediate categories 2, 3, and 4, the model appears to be somewhat less discriminating than it is for categories 1 and 5.

When evaluating the predictive power of models for polytomous dependent variables, it is important to keep in mind that what matters is not the proportion of dark and light lines in each plot (which will, of course, depend upon how the outcome variable is distributed between the possible categories), but rather the overall extent to which the dark lines are separated from the light lines. For example, in the case of the Neumayer model in Figure 13, many more country-year cases in that data set have Political Terror Scale values of 2, 3, and 4 than the extremes of 1 and 5.

Implementation in R

We have written an R package that allows the user to create separation plots. Details of this package are described in the online appendix that accompanies this article.

Conclusions

The separation plot allows one to assess the fit of a (logit or probit) regression model for which the dependent variable is binary. It adds information about the fit of discrete regression models and can enhance our ability to understand the uncertainty of the predictions which are made by such statistical models. The separation plot has the following characteristics:

- It provides a quick visual summary of the distribution of events and nonevents in the data. Thus, are the events rare or frequent?
- It illustrates whether observations with high predicted probabilities actually experienced the event. If so, the model has a high degree of predictive fit, which can be assessed visually, without predetermining the threshold for binning predictions.
- It illustrates the existence of clusters of false positives and false negatives, if either or both exist.
- It permits a fairly direct comparison between different models, both for the same data and for different data.
- It is relatively easy to implement, explain, and present. Since it is visual, it is grasped quickly.
- Just to be sure, it does not substitute for the use of simulation to compare expected values under different scenarios.

What is the downside of the separation plot? The basic downside is its upside. It is a visual presentation, one that has a certain subjective element to it. There are many single number summaries that could be employed to assess the fit of these kinds of models. Some of these are also arbitrary. What is the value for the area under the curve that indicates a model has a high degree of fit? Others, e.g., likelihood ratios, have a basis in statistical theory but generally require some comparison with other (often imaginary) models. The separation plot does not provide a single number summary, but rather provides a visual presentation of the fit. In this way it goes along with presentations of distributions of expected values under different scenarios, in which specific numerical tests are rarely presented and the extent of differences is portrayed graphically. However, it would be possible to calculate a variety of interesting statistics on the separation plot, including but not limited to the Mann-Whitney U test. We prefer to leave the development of this to scholars who may find them necessary. Rather, we think that the use of separation plots to present the fit of logistic and probit models will serve the needs of many social scientists.

References

Ahlquist, John S. 2010. “Building Strategic Capacity: The Political Underpinnings of Coordinated Wage Bargaining.” *American Political Science Review* 104(1): 171–88.

Aldrich, John, and Forrest Nelson. 1984. *Analysis with a Limited Dependent Variable: Linear Probability, Logit, and Probit Models*. Beverly Hills, CA: Sage.

Brier, Glenn W. 1950. “Verification of Forecasts Expressed in Terms of Probabilities.” *Bulletin of the American Meteorological Society* 78: 1–3.

Fearon, James D., and David D. Laitin. 2003. “Ethnicity, Insurgency, and Civil War.” *American Political Science Review* 97: 75–90.

Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. 2002. “Let’s Practice What We Preach: Turning Tables into Graphs.” *American Statistician* 60(4): 328–31.

Harrower, Mark, and Cynthia A. Brewer. 2003. “Colorbrewer .org: An Online Tool for Selecting Colour Schemes for Maps.” *The Cartographic Journal* 40(1): 27–37.

Herron, Michael C. 1999. “Postestimation Uncertainty in Limited Dependent Variable Models.” *Political Analysis* 8(1): 83–98.

Hillygus, D. Sunshine, and Simon Jackman. 2003. “Voter Decision Making in Election 2000: Campaign Effects, Partisan Activation, and the Clinton Legacy.” *American Journal of Political Science* 47(4): 583–96.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. “Toward a Common Framework for Statistical Analysis and Development.” *Journal of Computational and Graphical Statistics* 17(4): 892–913.

Kastellec, Jonathan P., and Eduardo L. Leoni. 2007. “Using Graphs Instead of Tables in Political Science.” *Perspectives on Politics* 5(4): 755–71.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. “Making the Most of Statistical Analyses: Improving Interpretation and Presentation.” *American Journal of Political Science* 44(2): 347–61.

Marshall, Monty G., and Keith Jaggers. 2003. *Political Regime Characteristics and Transitions 1800–2003*. Polity IV Project.

McKelvey, Richard D., and William Zavoina. 1975. “A Statistical Model for the Analysis of Ordinal Level Dependent Variables.” *Journal of Mathematical Sociology* 4(2): 103–20.

Neumayer, Erik. 2005. “Do International Human Rights Treaties Improve Respect for Human Rights?” *Journal of Conflict Resolution* 49(6): 925–53.

O’Brien, Sean. 2010. “Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research.” *International Studies Review* 12: 87–104.

Rosenstone, Steven J., and John M. Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Macmillan.

Sedgewick, Robert. 1998. *Algorithms in C*. Upper Saddle River, NJ: Addison-Wesley Professional.

Sedgewick, Robert, and Kevin Wayne. 2011. *Algorithms*. Upper Saddle River, NJ: Pearson Education.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15545–50.

Tufte, Edward R. 2006. *Beautiful Evidence*. Cheshire, CT: Graphics Press.

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010. “The Perils of Policy by P-Value: Predicting Civil Conflict.” *Journal of Peace Research* 47(4): 363–75.

Wood, Reed M., and Mark Gibney. 2010. “The Political Terror Scale (PTS): A Re-introduction and a Comparison to CIRI.” *Human Rights Quarterly* 32: 367–400.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix: R Package

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.