# Space and Some Other Things: Point Process Models for Hockey Data

*Devan Becker dbecker7@uwo.ca*

*April 14, 2017*

## 1 Introduction

Statistical methods for the analysis of baseball are well established. Each game is essentially a series of trials with a finite number of setups and outcomes. Each play begins the same and each player can only make one decision. After the play, all of the players are reset and another trial begins. Because of this, there are established statistical methods that can be directly applied. These methods have lead to a revolution in the study of baseball (e.g. James (1987)).

Hockey is not as amenable to such methods. There are not discrete plays that can be repeated and minor changes can lead to major consequences. A small mistake in a play setup can make the difference between a goal and an interception, which will cause different plays further down the line. This is an example of Chaos Theory in action, and thus analysis is not straightforward and requires careful consideration.

One primary difference between baseball and hockey is the importance of location. All pitchers are pitching from the same place, and all batters are batting from the same place. In hockey, the location of a play is crucial to its success. For this reason, we look at plays as points in space and time. This leads us naturally to the field of spatial point processes.

Point process models are used widely in the modelling of forest fires and earthquakes (see, e.g. Brillinger, Preisler, and Benoit (2003), Peng, Schoenberg, and Woods (2005), or Ogata (1998)). In these models, the location of ignition or the epicenter of the earthquake are treated as a single point. Forest fire data also includes other measured variables - such as duration, size, etc. - that are associated with each ignition. These are termed marks and are often related to specific scientific hypotheses. For a thorough review of point processes, see Daley and Vere-Jones (2003). For a complete treatment of spatio-temporal point processes, see P. J. Diggle (2013).

For point processes that only contain certain "types" of points, the marks can be regarded as labels. In the present research, we are concerned with determining whether points are randomly labelled (P. J. Diggle (2013)). That is, we want to know if the labels of each point are related to the location. The methods that we employ allow us great freedom in determining what we mean by labels. For this analysis, we have chosen four different definitions. These are as follows: period 1 and period 2; left/right side of the goalie; shots versus goals; and shots taken while winning versus losing. For each of these, we can determine the spatial association of shots within each label.

The data were scraped from NHL.com for the 2015-2016 season. For each game, the entire play-by-play summary was obtained. These summaries are separated into types of plays: hits, shots, goals, penalties, or fights. In total, there are 151,829 different plays. Of these plays, 71,683 were shots and 7,066 were goals (the number of shots does not include the number of goals). For each play, the location on the ice was recorded along with the time, period, team name, jersey number, and current goals for/against.

# 2 Methods and Necessary Preliminaries

## 2.1 Random Labelling

In our analyses, we wish to determine if there is a difference between the spatial distribution points with different labels. To do this, we consider the hypothesis of random labeling.

In a multi-type point process, we have locations of points and each point has a label assigned to it. The hypothesis of random labeling tests whether the labels assigned to those points were assigned randomly. In particular, it tests whether the assignments of the labels happened independently of the location of the labels. Following Lotwick and Silverman (1982), P. J. Diggle and Chetwynd (1991), and Gatrell, Bailey, Diggle, and Rowlingson (1996), we use a test for random labeling based on the inhomogeneous K-function.

There is a minor philosophical point that must be made here. We are testing whether the labels of the points are randomly assigned, but obviously the labels are assigned deterministically. However, the phrase "randomly assigned" may be a bit misleading. We are testing for an association between location and label, and a lack of association is interpreted as random labeling. In other words, we are testing to see if knowing the label gives us a better prediction for the location of the point (and vice versa). See Goreaud and Pélissier (2003) for more discussion.

## 2.2 K-Functions

The K-function is a measure of the average number of points within a given radius for any given point (Ripley (1979)). In other words, we look at every point in our dataset, draw a circle around it with a specific radius, and count the average number of points in each circle with that radius. We repeat this for many different radii. For homogeneous point processes, there will be the same number of points in each circle regardless of where the circle is centered, and the expected number of points will increase with the area of the circle. Formally, this is defined as:

$$K(r) = \frac{1}{\lambda} E(\text{number of points within distance } r \text{ of each point})$$

The parameter $\lambda$ is the intensity (or rate), and it is calculated as the number of points divided by the area. For a homogeneous point process, this parameter is a constant. For the inhomogeneous case, we modify the K-function slightly. The parameter $\lambda$ becomes a function, $\lambda(x)$, where $x$ is the location of an arbitrary point. For multi-type point processes, there are several K-functions. We can calculate the K-function only for points of type 1 and just points of type 2, but we can also examine the cross-K-functions, defined as follows:

$$K_{ij}(r) = E\left(\frac{\text{number of type } j \text{ points within radius } r \text{ of type } i \text{ points}}{\lambda(x_i)}\right)$$

Here, $\lambda(x_i)$ is the inhomogeneous intensity function, which, in essence, tells us how likely a point is at location $x_i$. Intuitively, the cross-K-function is the average number of type $j$ points near points of type $i$. If $i = j$, this reduces to the K-function for a single type.

Under the random labeling hypothesis, the type of point is chosen at random, and the location of the point is irrelevant (Baddeley and Turner (2005)). This means that it doesn't matter whether we're looking at points of type 1 or 2, we expect the K-function to be the equivalent. In particular, $K_{11}(r) = K_{22}(r) = K_{12}(r) = K(r)$, where $K(r)$ is the common K-function calculated without considering the labels.

Due to random variation, we don't expect any of the cross K-functions to be exactly equal. Instead, we need to define what we accept as approximately equal. This is where tests of statistical significance come in to play. In general, a significance test consists of finding a sample statistic (e.g. an average) and using the variance to define what hypotheses would be reasonable. The statistical test of the random labeling hypothesis is based

on the K-function. We cannot calculate the exact estimated variance of the K-function and, therefore, we use Monte Carlo Simulation.

### 2.3 Statistical Significance Through Random Sampling

Suppose you have a bottle cap and you want to know if it's fair (equally likely to land heads and tails). Suppose you flip it 20 times and you get 6 heads. If the coin is fair, how likely is this result?

One way to evaluate this is through Monte Carlo simulation. The basic idea is to simulate events under the assumption that the null hypothesis is true, i.e. the bottle cap has a 50% chance of being heads. We can simulate 20 flips of a fair coin and examine the result. If we do this 1,000 or 10,000 or 100,000 times, we can see what the probability of our sample is, given that the bottle cap has a 50% chance of heads.

In particular, we can see how likely our observed data is. We expect, under the null hypothesis, to see 10 heads. To see whether our coin is fair, we can look at the probability of seeing 6 or fewer heads. Similarly, we can look at the probability of seeing 14 or more heads. In other words, we're looking at the probability of being at least 4 away from 10 in either direction, given the null. This is the empirical p-value of a Monte Carlo test.

For 10,000,000 trials, I simulated 20 coin flips and recorded the number of heads. 11.499% of these simulations had 6 or fewer or 14 or higher heads. This means our p-value is 11.499%. Usually, a p-value of less than 0.05 is considered significant. Since our p-value is larger than 0.05, we would not be able to conclude that our bottle cap is unfair. We only saw 6 heads out of 20, but this isn't different enough from 10 out of 20 to be statistically significant.

### 2.4 Monte Carlo Estimation of the K-Function

In the example above, we could have used the binomial distribution to find the exact p-value (which is %11.531). We do not have this option with K-functions, so instead we must rely on Monte Carlo methods.

To test the hypothesis of random labeling we are still, essentially, looking at flipping coins. We assume that the locations of the points are fixed and the labels were assigned by coin flips. We randomly re-label each point, then we find the difference between the K-functions for each type, i.e. $D(r) = K_{11}(r) - K_{22}(r)$. We repeat this process, say, 100 times, and we get an estimate of what $D(r)$ should look like under the null hypothesis. These simulated functions act like a confidence interval. If the estimate of $D(r)$ for our data is outside of the simulated values, then we conclude that our points are not randomly labelled and thus the locations and the labels are somehow related.

Note that this does not reveal the locations in which the labels are different, it merely tests whether there are differences. If we conclude that the spatial distributions of labels are sufficiently different, we can further investigate what it is that makes them different.

In summary, we are estimating $D(r)$, the difference in K-functions for different labels, and comparing it to estimates found by randomly relabeling the points. If the estimated $D(r)$ is outside of the 95% confidence bands created by the simulation, then we can conclude that the labels have a different spatial distribution.

## 3 Results

Thankfully, we return to hockey. The four hypotheses mentioned in the introduction are tested using the methods outlined above.

### 3.1 Shots in Period 1 versus Period 2

Goalies switch sides between periods, but the benches don't. This means that players will have further to travel when coming off the bench. The last couple minutes of period 3 are notably different from the rest of the game, so we focus on periods 1 and 2. The rest of this paper will also be restricted to periods 1 and 2 in order to limit the number of unusual situations (e.g. pulled goalies, last ditch efforts).

As can be seen in the plot on the left of Figure 1, it does not appear that the period affects the location of the shots. Since the estimated function $D(r)$ falls within the range of 95% of the randomly simulated data, we conclude that the period number and location of shots are not dependent. That is, players shoot from similar places in both periods.

### 3.2 Shots While Losing versus Winning

When a team is losing, it is reasonable to suspect that they simply take any shot that they can. There's a certain amount of randomness for each goal, and probability favours the bold. Again, our intuition betrays us. The estimated $D(r)$ is well within the Monte Carlo envelope (right hand side of Figure 1), so we conclude that the locations of the shots do not appear to depend on whether the team is winning or losing.

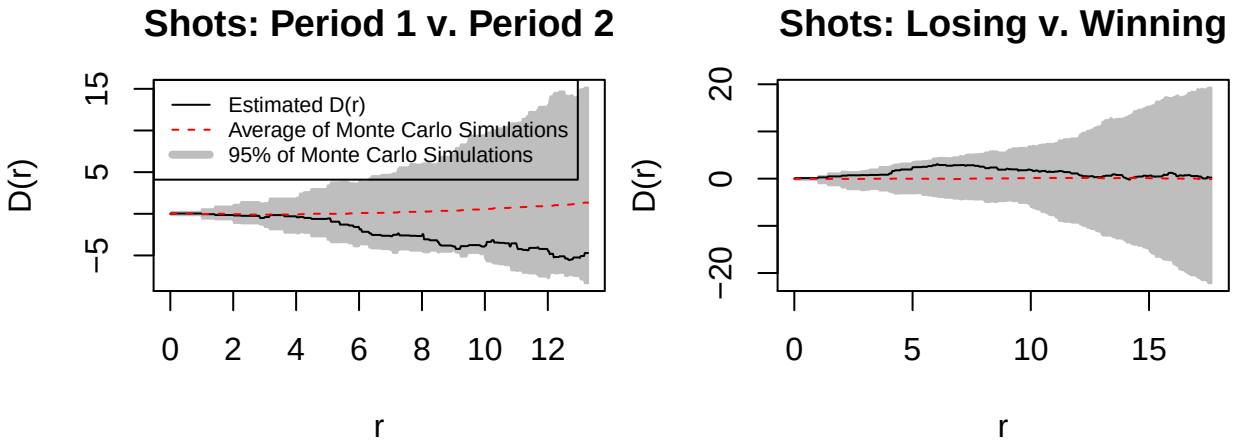


Figure 1: Difference in K-Functions for period 1 vs 2 (left) and losing vs winning (right). . If the estimated D(r) (black) falls within 95% of the simulated values (shaded region), then the true data is indistinguishable from the simulations.

### 3.3 Net Symmetry

The NHL does not have an equal distribution of left- and right-handed players. Furthermore, the goalies themselves are not symmetric. The aim of this analysis is to determine whether the shots taken on the left side of the net are symmetric to the right side. It is worth noting that, in the data used for this particular analysis, there were 3,427 shots taken from the left side of the net and 3,639 from the right side.

To test this hypothesis, we must transform the data. In this transformation, we first split it up into left and right (Figure 2, top left). Then, we fold the data back on itself so that all of the shots are moved to the upper-right quadrant. The middle plot in the top row of Figure 2 is the exact same data as the top left plot. This allows the left and right of the goalie to be directly comparable.

Now that the data is all defined on the same region, we can apply our test. The resulting plot is the bottom row of Figure 2.

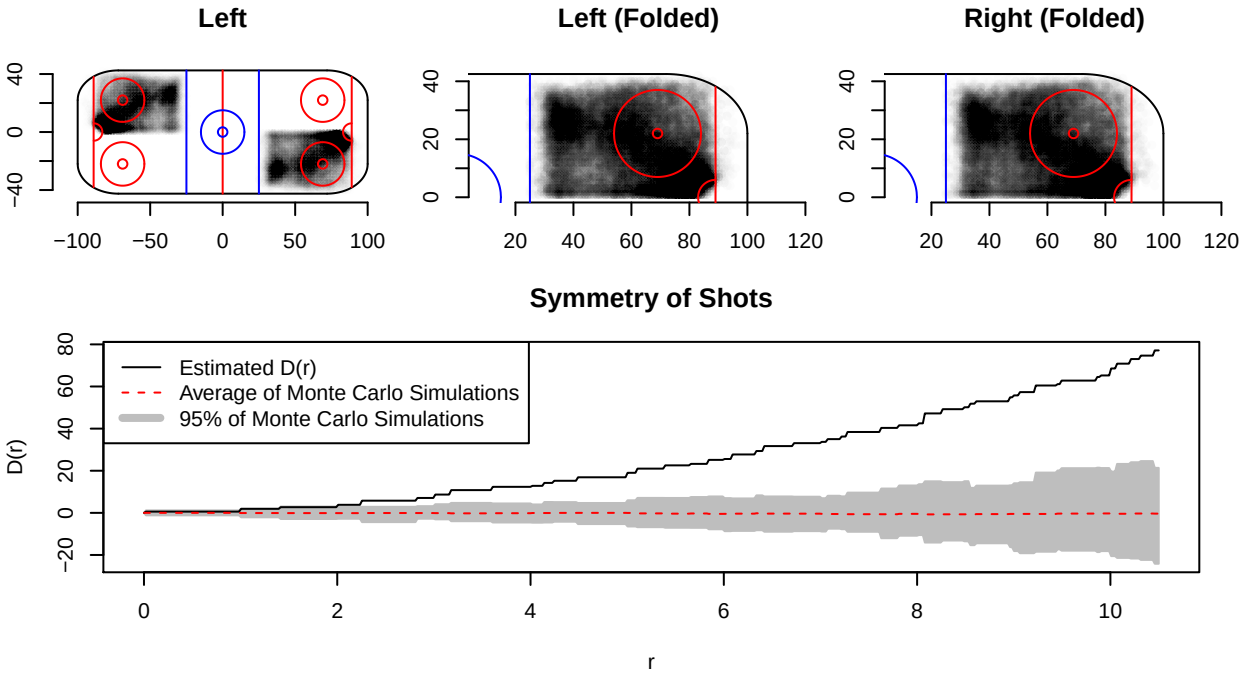


Figure 2: Testing for the symmetry of shots. The top row is the data transformation used. The bottom plot shows the estimate of $D(r)$ along with the 95% Monte Carlo simulation envelope.

### 3.4 Shots Versus Goals

The final analysis in this series is focused on the difference between shots and goals. Again, we restrict our data to the first and second periods because the third period is systematically different.

The results of this are shown in Figure 3. The estimated function $D(r)$ is different from the 95% confidence bands found by the Monte Carlo simulation, thus we concluded that goals happen in different locations than shots that were not goals.

## 4 Conclusions and Future Work

In this paper, we have explored some of the relationships between locations of shots and various labels for those shots. These methods can lead to some interesting analyses, and we hope that our results will motivate further research in this area.

The period does not appear to effect the location of shots on a large scale. This is somewhat surprising, but it is pretty rare that a player will take advantage of (or be disadvantaged by) the distance needed for a substitution.

Surprisingly, the score does not effect where the players shoot from. We expected the players to make more desperate plays when losing and more cautious plays while winning, but the data does not seem to support that assumption.

Unsurprisingly, shots are not symmetric around the goalie. Based on this result, we can inspect the intensity functions to see where the two sides differ. An early approximation of this is shown in the Appendix. It should be noted that the estimate of the intensity function is not accurate at the edges of the plot. However, it does appear that the goalie gets more close shots on his left hand side.

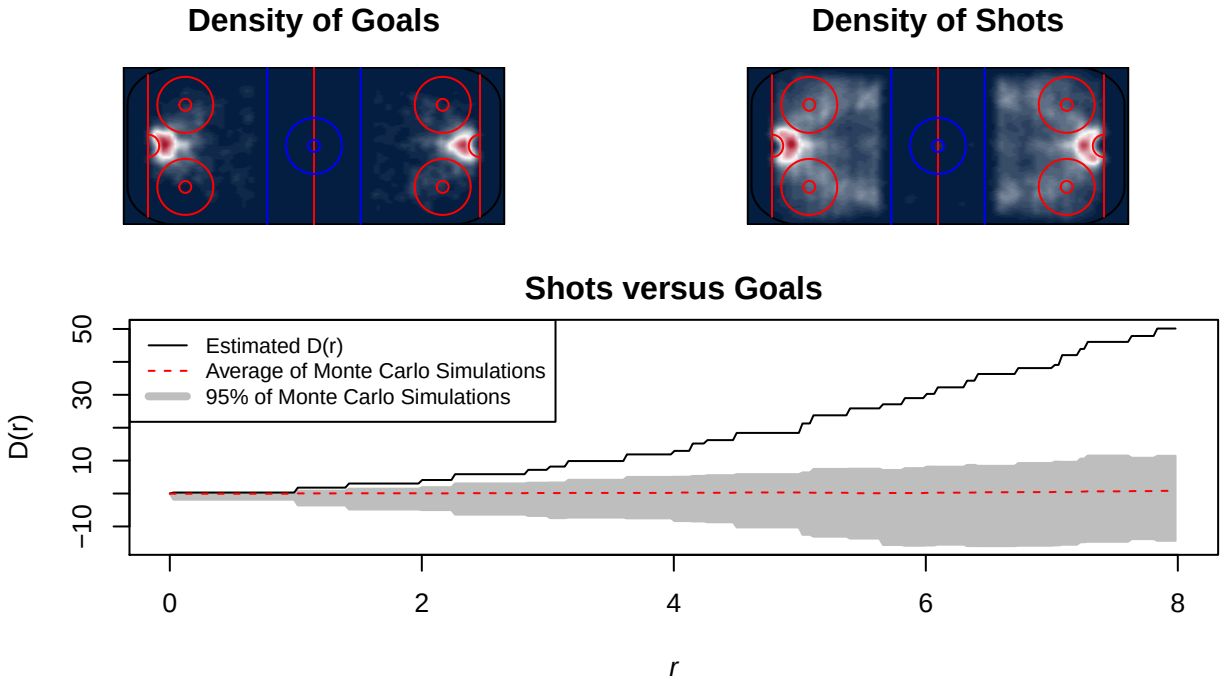


Figure 3: The top row of plots shows the estimated intensity of goals and shots. Red values indicate more shots taken from that area. Note that our definition of shots does not include goals. The plot on the bottom row is an estimate of $D(r)$ for shots versus goals.

Also unsurprising is the result that shots and goals occur at different locations. This demonstrates that not all shots have the same chance of scoring. This also demonstrates that not all shots are meant to score. Many shots are taken simply to put pressure on the goalie or to force a rebound. This gets the puck closer to the net, and our analysis, as well as logic, show that shots taken closer to the net are more likely to be goals.

We note that the labels in this paper could also be seen as temporal covariates. A mark is a value that can only be measured at a point, while a covariate is a value that can be measured anywhere. Since we treated the data as spatial, aggregated by period, we believe that the use of marks is justified.

We treated the data as spatial for our purposes, but many of the analyses could be improved by incorporating a temporal component. The periods in a hockey game will complicate any temporal analysis, but these challenges can be overcome.

Notably missing from this paper is the study of power plays. It is our belief that power plays would not affect the results that were presented here, but in future analysis incorporating power plays is recommended.

For the most part, this study was exploratory. We have demonstrated the existence of general trends and found purely descriptive statistics. Further research is needed before we reach any prescriptive conclusions. Specifically, we need to restrict the scope of our analyses to specific situations, such as the shots versus goals when on a breakaway or at full strength. The present dataset may not be sufficient, but many of the recorded values are easily cross-referenced.

## Appendix: Estimates of the Difference

The plots in Figure 4 show the difference in estimates for the labels that were found to be significantly different. Blue values represent low values and red values represent high. The plot on the left indicates that

there may be more shots on the left had side of the goalie than the right. There is also an area with a deep blue directly in front of the net. This may be an artifact of the estimation or it may be that shots are more likely to occur slightly to the right in front of the goalie. Further research is needed.

The plot on the right has deeper values of blue further away from the net, and bright values of red closer to the net. This indicates that there are many shots taken from the point that do not go in, and most of the shots that do go in are taken from directly in front of the goalie. Further study might investigate situations in which this is not true, i.e. when shots from the point are more likely to go in.

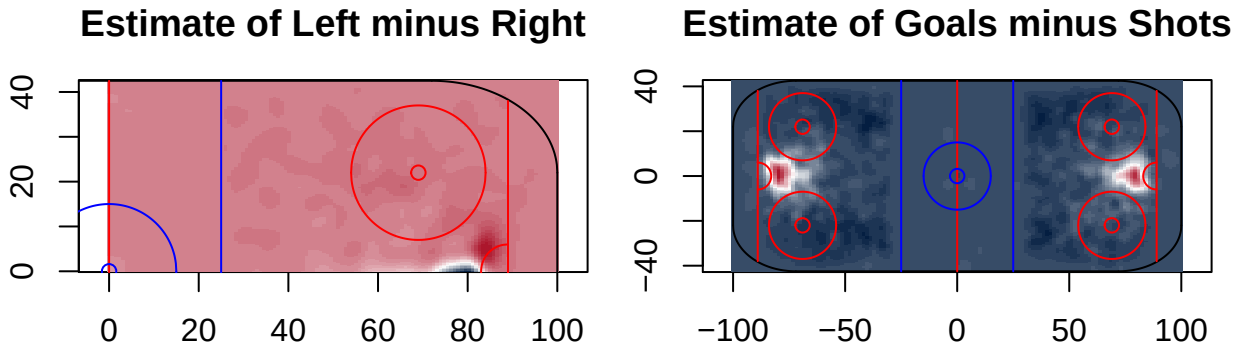


Figure 4: (Left) Difference in the estimates of the intensity function for the left and right side of the goalie. (Right) Differences in the estimates of Goals and Shots.

## 5 Acknowledgements

The author would like to thank Jonathon Gascho and Erin Lundy for their contributions to this research and help with the writing process. The author would also like to acknowledge his supervisors, Douglas Woolford, W. John Braun, and Charmaine Dean, for their guidance in his research in this area of statistics.

## References

Baddeley, A., and Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal Of Statistical Software*, *12*(6), 1–42.

Brillinger, D. R., Preisler, H. K., and Benoit, J. W. (2003). Risk assessment: a forest fire example. *Lecture Notes-Monograph Series*, *40*(2003), 177–196.

Daley, D., and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press.

Diggle, P. J., and Chetwynd, A. G. (1991). Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations. *Biometrics*, *47*(3), 1155–1163.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Trans Inst Br Geogr*, *21*(1), 256–274. https://doi.org/10.2307/622936

Goreaud, F., and Pélissier, R. (2003). Avoiding Misinterpretation of Biotic Interactions with the Intertype K12-Function : Population Independence vs . Random Labelling Hypotheses. *Journal of Vegetation Science*,

*14*(5), 681–692.

James, B. (1987). *The Bill James Baseball Abstract 1987*. Ballantine Books.

Lotwick, H., and Silverman, B. W. (1982). Methods for Analysing Spatial Processes of Several Types of Points. *Journal of the Royal Statistical Society - Series B (Statistical Methodology)*, *44*(3), 406–413.

Ogata, Y. (1998). Space-time point-process models for earthquake occurences. *Annals of the Institute of Statistical Mathematics*, *50*(2), 379–402.

Peng, R. D., Schoenberg, F. P., and Woods, J. A. (2005). A Space–Time Conditional Intensity Model for Evaluating a Wildfire Hazard Index. *Journal of the American Statistical Association*, *100*(469), 26–35. https://doi.org/10.1198/016214504000001763

Ripley, B. D. (1979). Tests of "Randomness" for Spatial Point Patterns. *Journal of the Royal Statistical Society - Series B (Statistical Methodology)*, *41*(3), 368–374. https://doi.org/10.2307/2346101