Using machine learning to identify structural breaks in single-group interrupted time series designs

Ariel Linden, DrPH[1,2], Paul R. Yarnold, PhD[3]

[1] President, Linden Consulting Group, LLC - Ann Arbor, MI alinden@lindenconsulting.org

[2] Research Scientist, Division of General Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

[3] President, Optimal Data Analysis, LLC

Corresponding Author Information:
Ariel Linden, DrPH
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI USA 48103
Phone: (971) 409-3505
Email: alinden@lindenconsulting.org

Key Words: interrupted time series analysis, quasi-experimental, causal inference, structural breaks, machine learning, data mining, optimal discriminant analysis, maximum-accuracy model

Running Header: machine learning to identify structural breaks

Acknowledgement: We wish to thank Julia Adler-Milstein for reviewing the manuscript and providing many helpful comments.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jep.12544

ABSTRACT

Rationale, aims and objectives: Single-group interrupted time series analysis (ITSA) is a popular evaluation methodology in which a single unit of observation is being studied, the outcome variable is serially ordered as a time series, and the intervention is expected to "interrupt" the level and/or trend of the time series, subsequent to its introduction. Given that the internal validity of the design rests on the premise that the interruption in the time series is associated with the introduction of the treatment, treatment effects may seem less plausible if a parallel trend already exists in the time series prior to the actual intervention. Thus, sensitivity analyses should focus on detecting structural breaks in the time series before the intervention.

Method: In this paper, we introduce a machine-learning algorithm called optimal discriminant analysis (ODA) as an approach to determine if structural breaks can be identified in years prior to the initiation of the intervention, using data from California's 1988 voter-initiated Proposition 99 to reduce smoking rates.

Results: The ODA analysis indicates that numerous structural breaks occurred prior to the actual initiation of Proposition 99 in 1989, including perfect structural breaks in 1983 and 1985, thereby casting doubt on the validity of treatment effects estimated for the actual intervention when using a single-group ITSA design.

Conclusions: Given the widespread use of ITSA for evaluating observational data and the increasing use of machine-learning techniques in traditional research, we recommend that structural break sensitivity analysis is routinely incorporated in all research using the single-group ITSA design.

1. INTRODUCTION

Interrupted time-series analysis (ITSA) is a popular evaluation methodology for study designs in which a single unit of observation (e.g., an individual, a city, or a country) is being studied, the outcome variable is a serially ordered time series, and multiple observations are captured in both the pre- and post-intervention periods [1,2]. The study design is called an *interrupted time series* because the intervention is expected to "interrupt" the level and/or trend of the time series, subsequent to its introduction [3,4]. ITSA has strong internal validity, even in the absence of a comparison group, due primarily to its control over the effects of *regression to the mean* [3,5,6]. When the treatment group's outcomes can also be contrasted with those of one or more comparison groups, the internal validity is further enhanced by allowing the researcher to potentially control for confounding omitted variables [2]. Additionally, ITSA has strong external validity when the unit of measure is at the population level, or when the results can be generalized to other units, treatments or settings [4,7].

ITSA has been used in many areas of study, such as assessing the effects of community interventions [8,9], public policy [10], regulatory actions [11], and health technology assessment [12], to cite but a few. ITSA has also been proposed as a more flexible and rapid design to be considered in health research before defaulting to the traditional two-arm randomized controlled trial [13]. In addition, systematic reviews of the literature increasingly include studies using ITSA as the primary research design [14].

The validity of ITSA when used for making causal inferences has begun to receive attention in the literature, specifically the importance of testing for interruptions in the time

series that occur prior to the actual initiation of the intervention [2]. The assumptions necessary for causal inference in the single-group ITSA may seem plausible when the pre-intervention trend is followed by a statistically significant change in the trend of the outcome variable immediately following the introduction of the intervention, and sustained over some meaningful period of time. In contrast, these assumptions seem less plausible if a parallel trend already exists in the time series prior to the initiation of the intervention. Linden [2] suggests conducting an iterative sensitivity analysis involving testing each pre-intervention time period treated as a "pseudo-intervention" period. This approach is consistent with regression-based *structural break* analysis commonly used in time series econometrics (Hansen [15] and Perron [16] provide excellent reviews of structural break analysis literature). The underlying assumptions of the single-group ITSA may be challenged if interruptions in the level or trend of the outcome variable are found to exist at other time points prior to the actual initiation of the intervention.

This paper introduces a machine learning algorithm called optimal discriminant analysis (ODA) [17] to determine if (and to what degree) structural breaks can be identified in periods prior to the actual initiation of the intervention. This methodology has several noteworthy strengths in that it provides intuitive measures of predictive accuracy (e.g., sensitivity, specificity, effect strength for sensitivity), model-free permutation tests to derive *P* values, and cross-validation to assess generalizability of the model to new cases applied in similar settings [7,18,19,20,21,22,23,24]. It is therefore likely to be an approach that will be of interest to those using ITSA designs as well as those more generally interested in applications of machine learning to traditional research designs. To illustrate the ODA approach for assessing

interruptions in times series data prior to the actual initiation of the intervention, Section 2 describes study background, data, ODA methodology, and analytic strategy; Section 3 reports the findings; and Section 4 presents discussion and conclusions.

2. METHODS

2.1 Background and data

We examine data from the 1988 voter-initiated Proposition 99, a widespread effort in California to reduce smoking rates by raising the cigarette excise tax by 25 cents per pack, and to fund anti-smoking campaigns and other related activities throughout the state (Breslow & Johnson [25] provide a comprehensive discussion of this initiative).

Per capita cigarette sales (in packs) is the most widely used indicator of smoking prevalence found in the tobacco research literature [26] and serves here as the aggregate outcome variable under study, measured annually at the state level from 1970 until 2000 (with 1989 representing the first year of the intervention). The current data were obtained from Abadie et al. [27], who obtained the data from Orzechowski & Walker [28]. The current study limits analysis to cigarette sales in only the pre-intervention years between 1970 and 1988 to determine if there are additional interruptions in the time series prior to actual initiation of the intervention in 1989.

2.2 Brief description of Optimal Discriminant Analysis (ODA)

ODA is a machine learning algorithm introduced over a quarter century ago [17]. Derived using mathematical programming methods, ODA was developed as a methodology for identifying exact non-parametric statistical models that *explicitly maximize predictive accuracy* normed

against chance [23,24]. The objective function maximized by ODA—predictive accuracy—is in contrast to alternative methods developed to explicitly maximize the amount of variance explained, or the value of the likelihood function [29,30]. In general, for an ordered or continuous variable (e.g., an outcome score), and a two-category class variable (e.g., an intervention), an ODA model has the form: if score ≤ (value) predict that the observation is from class A; otherwise predict that the observation is from class B. ODA identifies the cut-point that explicitly maximizes the predictive accuracy of the model (i.e., in terms of the correct classification of actual members of class A and of class B) indexed using the *effect strength for sensitivity* (ESS) statistic described below.

2.2.1 Assessing statistical significance of ODA models

Statistical significance (*P* value) of ODA models is computed as a permutation probability: no distributional assumptions are required of the data and *P* values are exact [17,24]. In study designs involving two or more tests of statistical significance, a sequentially-rejective Sidak Bonferroni-type multiple comparisons methodology is used to prevent "alpha inflation" and ensure the desired experimentwise *P* value (here, $P < 0.05$) [23].

2.2.2 Ecological significance of ODA models

Ecological significance (normed accuracy) of ODA models is assessed using the ESS statistic -- a *chance*-corrected (0 = the level of predictive accuracy expected by chance) and *maximum*-corrected (100 = perfect prediction) index of the predictive accuracy of a statistical model (computation of ESS is discussed elsewhere [19,20,22;23,24]). The cut-point identified by ODA *explicitly maximizes* the ESS yielded by the ODA model developed for the total ("training")

sample. Using ESS, investigators may directly compare the predictive accuracy of different models (relative to chance), regardless of structural features such as sample size, skew, or "outliers" [24]. By convention, ESS values of 25% or less indicate a relatively weak effect, values of 50% or less indicate a moderate effect, values of 75% or less indicate a relatively strong effect, values of 90% or less indicate a strong effect; and ESS values greater than 90% indicate a very strong effect [23].

2.2.3 Assessing generalizability of ODA models

Cross-validation in the ITSA context connotes estimating the generalizability of the model when it is applied to future points in the time series, or to similar series (e.g., other states implementing anti-smoking campaigns) assuming they are comparable on other characteristics. Several algorithms commonly used to estimate model generalizability include *k*-fold cross-validation, bootstrapping, and leave-one-out jackknife (LOO) cross-validation [19,23,24,31,32]. Presently ODA implements the (LOO) approach, which is simply *n*-fold cross-validation, where *n* is the number of observations in the dataset. Each observation in turn is left out, the predicted class membership is obtained for the hold-out observation, and accuracy is determined as success or failure in predicting the actual class membership of that observation. The results of all *n* predictions are used to calculate LOO (validity) accuracy, which is then compared to total sample (training) accuracy.

2.3 Analytic approach

While there currently is no "rule of thumb" for defining the circumstances in which structural breaks invalidate treatment effect estimates drawn from an ITSA model, the identification of one

or more structural breaks in the pre-intervention period should serve as an indicator that further scrutiny of the data is warranted, and assumptions of a treatment effect should be challenged. In order to systematically assess the presence or absence of structural breaks in years prior to 1989, a series of eighteen "pseudo-interventions" was generated -- one for each year commencing with 1970 and ending with 1988. For example, in the pseudo-intervention year 1970, the intervention is set to 1 for all years from 1971 onward, while 1970 represents the sole pre-intervention period and is set to 0. At the other end of the continuum, this layout is reversed, with the final pseudo-intervention year being 1988: here, all years from 1970 to 1987 represent pre-pseudo-intervention periods and are set to 0, while 1988 represents the pseudo-intervention year and is set to 1.

In ODA each pseudo-intervention is treated as a class variable with two categories -- either pre- (0) or post- (1) pseudo-intervention period. In this study the relationship between the pseudo-intervention and per capita cigarette pack sales was ascertained using an ODA model of the form: if annual cigarette sales $\leq$ (cut-point) then predict that the observation is from the post–pseudo-intervention; otherwise predict the observation is from the pre-pseudo-intervention period (in the actual analysis a non-directional "two-tailed" hypothesis was tested for all ODA analyses).

As this study involved a total of 18 tests of statistical significance, we controlled for the effect of multiple testing by performing a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure to ensure an experimentwise $P < 0.05$ [23]. Exact P values were estimated using 25,000 Monte Carlo experiments.

Finally, the upper-bound of expected cross-generalizability of ODA models across time was assessed using LOO analysis, in which sequential classification of every year in the training sample was used to generate a model holding out the year being classified. When identical ESS is obtained in training and validity analysis this suggests the training model may cross-generalize to the following year with comparable reliability and strength. However, obtaining ESS that is lower in LOO than in training analysis suggests the cut-point that maximizes predictive accuracy in training analysis may not cross-generalize to the following year with comparable reliability and strength [23,24]. LOO requires at least two observations in both the pseudo-pre and post-intervention periods, and thus is not reported for pseudo-interventions with one observation per class category (1970, 1987, and 1988).

3. RESULTS

The Table presents the annual actual cigarette sales per capita, the ODA derived cutpoint on cigarette sales for predicting belonging to the pre- and post-pseudo-intervention periods, and reliability and accuracy measures (*P* values and ESS) for training and LOO analysis. While no ODA model could be obtained for 1988, and no LOO model could be obtained for 1970, 1987, or 1988, ODA identified statistically significant structural breaks (i.e. generalized $P < 0.05$) for all years between 1975 and 1986 based on analyses involving the total sample (training analysis), and between 1976 and 1985 when considering LOO cross-validation. When considering only structural breaks meeting the more stringent Sidak adjusted *P* values, then all years between 1977 and 1985 meet the experimentwise criterion for the training analysis, and 1979 through

1983 and 1985 met the experimentwise criterion in LOO cross-validation analysis. ESS values ranged from 53% to 100% for the training analysis (representing relatively strong to perfect effect strength), and from 28% to 100% for LOO analysis (representing moderate to perfect effect strength).

When considering Sidak-adjusted *P* values, ESS, and type of analysis (training *and* LOO) together, *perfect* structural breaks (i.e., the ESS in training and in LOO analysis are both 100%, and have experimentwise $P < 0.05$) are identified for the years 1983 and 1985, and strong, reproducible, statistically reliable structural breaks are identified for the years 1979 through 1982.

4. DISCUSSION

The present ODA analysis indicates that numerous structural breaks occurred prior to the actual initiation of Proposition 99 in 1989 -- including perfect structural breaks (i.e., ESS=100 in both training and LOO analyses) in 1983 and 1985 -- thereby casting doubt on the validity of treatment effects estimated for the actual intervention when using a single-group ITSA design [2]. More broadly, these results highlight the importance of routinely performing structural break analyses when using the single-group ITSA framework as a way to test the sensitivity of treatment effect estimates [33].

The ODA-based approach described here provides a robust framework for analyzing structural breaks in ITSA designs due to the following features. First, as a machine learning algorithm, ODA is not as constrained by a small number of observations as are conventional

statistics-based methods. This is particularly important for short time-series where regression-based structural break analyses fail to obtain parameter estimates for observations near to the beginning or to the end of the sample [34]. Of course, for ODA, the smaller a sample becomes, the greater the model ESS is needed to render a statistically reliable result, and for very small samples, only models achieving perfect or nearly perfect accuracy yield $P < 0.05$ [24,35,36].

Second, the ODA algorithm, with its associated measure of classification performance (ESS) and non-parametric permutation tests, can be universally applied to any variable type, and is not affected by skewed data or outliers -- a concern that may arise in the context of meeting assumptions underlying the validity of the estimated *P* value using conventional statistics [29,30]. And third, ODA can directly estimate the generalizability of the model when it is applied to future points in the time series, or to other interventions with similar characteristics.

In summary, for applications using the single-group ITSA framework for estimating treatment effects, this paper highlights the importance of -- and an intuitive, transparent machine learning methodology for -- assessing the existence of structural breaks that may occur in the time series prior to the initiation of an actual intervention. We recommend that structural break sensitivity analysis is routinely incorporated in all research using the single-group ITSA design, as a means of evaluating the unique efficacy of the actual intervention in influencing the trajectory of a temporal outcome measure.

REFERENCES

1. Linden, A., & Adams, J. L. (2011) Applying a propensity-score based weighting model to interrupted time series data: Improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice,* 17, 1231–1238.
2. Linden, A. (2015a) Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal,* 15, 480–500.
3. Campbell, D. T., & Stanley, J. C. (1966) *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.
4. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
5. Linden, A. (2007) Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes,* 15, 7-12.
6. Linden, A. (2013) Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology,* 13, 1-7.
7. Linden, A., Adams, J., & Roberts, N. (2004) The generalizability of disease management program results: getting from here to there. *Managed Care Interface,* 17, 38-45.
8. Biglan, A., Ary, D., & Wagenaar, A. C. (2000) The value of interrupted time-series experiments for community intervention research. *Prevention Science,* 1, 31–49.
9. Gillings, D., Makuc, D., & Siegel, E. (1981) Analysis of interrupted time series mortality trends: An example to evaluate regionalized perinatal care. *American Journal of Public Health,* 71, 38–46.

10. Muller, A. (2004) Florida's motorcycle helmet law repeal and fatality rates. *American Journal of Public Health*, 94, 556–558.
11. Briesacher, B. A., Soumerai, S. B., Zhang, F., Toh, S., Andrade, S. E., Wagner, J. L., Shoaibi, A., & Gurwitz, J. H. (2013) A critical review of methods to evaluate the impact of FDA regulatory actions. *Pharmacoepidemiology and Drug Safety*, 22, 986– 994.
12. Ramsay, C. R., Matowe, L., Grilli, R., Grimshaw, J. M., & Thomas, R. E. (2003) Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care*, 19, 613–623.
13. Riley, W. T., Glasgow, R. E., Etheredge, L., & Abernethy, A. P. (2013) Rapid, responsive, relevant (R3) research: A call for a rapid learning health research enterprise. *Clinical and Translational Medicine*, 2, 1–6.
14. Effective Practice and Organisation of Care (EPOC). (2015) Interrupted time series (ITS) analyses. EPOC Resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services. Available at: http://epoc.cochrane.org/epoc-specific-resources-review-authors
15. Hansen, B. E. (2001) The new econometrics of structural change: Dating breaks in US labor productivity. *The Journal of Economic Perspectives*, 15, 117-128.
16. Perron, P. (2006) Dealing with structural breaks. In Palgrave Handbook of Econometrics: Econometric Theory, Vol I, ed. T. C. Mills and K. Patterson, 278–352. Basingstoke, UK: Palgrave.

17. Yarnold, P.R., & Soltysik, R.C. (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, *22*, 739-752.
18. Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12, 132-139.
19. Linden, A., Yarnold, P. R. (In Print_A) Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*
20. Linden, A., Yarnold, P. R. (In Print_B) Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*
21. Linden, A. (2015b) LOOCLASS: Stata module for generating classification statistics of Leave-One-Out cross-validation for binary outcomes. Statistical Software Components s458032, Boston College Department of Economics. Downloadable from http://ideas.repec.org/c/boc/bocode/s458032.html [Accessed on 26 February 2016].
22. Linden, A. (2015c) CLASSTABI: Stata module for generating classification statistics and table using summarized data. Statistical Software Components s458127, Boston College Department of Economics. Downloadable from https://ideas.repec.org/c/boc/bocode/s458127.html [Accessed on 26 February 2016]
23. Yarnold, P.R., & Soltysik, R.C. (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.
24. Yarnold, P.R., & Soltysik, R.C. (2016) *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

25. Breslow, L., & Johnson, M. (1993) California's Proposition 99 on Tobacco, and Its Impact. *Annual Review of Public Health*, 14, 585–604.
26. Abadie, A., Diamond, A., & Hainmueller, J. (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105, 493–505.
27. Abadie, A., Diamond, A., & Hainmueller, J. (2014) SYNTH: Stata module to implement synthetic control methods for comparative case studies. Statistical Software Components S457334, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457334.html
28. Orzechowski, W., & Walker, R. C. (2005) *The Tax Burden on Tobacco. Historical Compilation*, vol. 40. Arlington, VA: Orzechowski & Walker.
29. Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, D.C.: APA Books, 1995.
30. Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: APA Books, 2000.
31. Witten, I.H., Frank, E., & Hall, M.A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*. San Francisco: Morgan Kaufmann.
32. Linden, A., Adams, J., & Roberts, N. (2005) Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes*, 13, 159-167.

33. Linden, A., Adams, J., & Roberts, N. (2006) Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice*, 12, 140-147.

34. Andrews, D. W. K. (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61, 821–856.

35. Yarnold, P.R. (2013). Percent oil-based energy consumption and average percent GDP growth: A small sample UniODA analysis. *Optimal Data Analysis, 2*, 60-61.

36. Yarnold, P.R. (2015). UniODA *vs*. McNemar's test: A small sample analysis. *Optimal Data Analysis*, *4*, 27-28.

Table. Cigarette sales per capita, ODA model cutpoint on cigarette sales for predicting group assignment in the pseudo-intervention period, and accuracy measures (*P* values and ESS) for corresponding training and LOO analyses.

Year	Per capital sales (in packs)	Predict intervention if sales <=	*Training Set*		*LOO Analysis*	
			P value	ESS	*P* value	ESS
1970	123.00	122.45	0.8474	61.11%	---	---
1971	121.00	120.60	0.5262	52.94%	0.3217	47.06%
1972	123.50	120.60	0.2956	56.25%	0.1703	50.00%
1973	124.40	120.60	0.1530	60.00%	0.3329	28.33%
1974	126.70	120.60	0.0686	69.23%	0.1842	37.14%
1975	127.10	120.60	0.0204*	75.00%	0.0914	44.87%
1976	128.00	120.60	0.0069*	81.82%	0.0399*	52.38%
1977	126.40	120.60	0.0018**	90.00%	0.0149*	60.23%
1978	126.10	120.60	0.0003**	90.91%	0.0045*	68.89%
1979	121.90	120.60	0.0001**	100.00%	0.0010**	78.89%
1980	120.20	119.40	0.0001**	100.00%	0.0002**	87.50%
1981	118.60	117.00	0.0001**	100.00%	0.0003**	85.71%
1982	115.40	113.10	0.0001**	100.00%	0.0006**	83.33%
1983	110.80	107.80	0.0003**	100.00%	0.0001**	100.00%
1984	104.80	103.80	0.0007**	100.00%	0.0158*	68.33%
1985	102.80	101.25	0.0020**	100.00%	0.0010**	100.00%
1986	99.70	98.60	0.0122*	100.00%	0.2047	44.12%
1987	97.50	93.80	0.1053	100.00%	---	---
1988	90.10	---	---	---	---	---

Notes:
--- No ODA model possible
** Experimentwise $P < 0.05$; * Generalized $P < 0.05$
ESS = effect size sensitivity (0 = chance accuracy, 100 = perfect accuracy)
LOO = leave-one-out (jackknife) cross-validation