Intern. J. of Research in Marketing 32 (2015) 94–112

Contents lists available at ScienceDirect

Intern. J. of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

Improving the statistical performance of tracking studies based on repeated cross-sections with primary dynamic factor analysis

Rex Yuxing Du [a,*], Wagner A. Kamakura [b,1]

[a] Bauer College of Business, University of Houston, Houston, TX 77204-6021, United States
[b] Jones Graduate School of Business, Rice University, 6100 Main Street (MS 531), Houston, TX 77005, United States

ARTICLE INFO

Article history:
First received on January 31, 2014 and was under review for 6 months
Available online 15 November 2014

Area Editor: Sandy D. Jap

Keywords:
Tracking study
Repeated cross-sectional survey
State-space model
Dynamic factor analysis

ABSTRACT

Tracking studies are prevalent in marketing research and virtually all the other social sciences. These studies are predominantly implemented via repeated independent, non-overlapping samples, which are much less costly than recruiting and maintaining a longitudinal panel that track the same sample over time. In the existing literature, data from repeated cross-sectional samples are analyzed either independently for each time period, or longitudinally by focusing on the dynamics of the aggregate measures (e.g., sample averages). In this study, we propose a multivariate state-space model that can be applied directly to the individual-level data from each of the independent samples, simultaneously taking advantage of three patterns embedded in the data: a) inter-temporal dependence within the population means of each variable, b) temporal co-movements across the population means of different variables and c) cross-sectional co-variation across individual responses within each sample. We illustrate our proposed model with two applications, demonstrating the benefits of making full use of all the available data. In the first illustration, we have access to all the individual-level purchase data from one large population of grocery shoppers over a span of 36 months. This provides us a testing ground for benchmarking our proposed model against existing approaches in a Monte Carlo experiment, where we show that our model outperforms all the alternatives in inferring population dynamics using data sampled through repeated cross-sections. We find that, as compared with using simple sample averages, our proposed model can improve the accuracy of repeated cross-sectional tracking studies by double digits, without incurring any additional data-gathering costs (or equivalently, reducing the data-gathering costs by double digits while maintaining the desired accuracy level). In the second illustration, we apply the proposed model to repeated cross-sectional surveys that track customer perceptions and satisfaction for an automotive dealer, a situation often encountered by marketing researchers.

1. Introduction

Tracking studies play an important role in monitoring population dynamics for various social, political, economic and business purposes. These studies typically rely on two basic sampling schemes: a) longitudinal panels, where data are gathered from the same sample of individuals over time, and b) repeated cross-sections, where data are gathered from different, independent samples in each period. Although both sampling schemes can be used to track population-level dynamics, only longitudinal panels can capture individual-level dynamics (e.g., within-individual attitude or behavior changes). However, because of the heavier burden on the participants in a longitudinal panel and the resulting challenges in recruiting and retaining panel members, maintaining representative longitudinal panels over an extended period of time is much more costly than recruiting repeated cross-sections (Hsiao, 2007). Indeed, the extra costs of longitudinal panels are in many cases unjustifiable because "relatively few analyses are truly longitudinal" (Tourangeau, 2003, p. 7) in the strict sense of studying individual-level dynamics.

Besides being less costly, repeated cross-sections can often provide a better representation of changing populations than do longitudinal panels. For example, with highly mobile populations such as the younger and less affluent, it is often difficult to track panel members as they move, further eroding sample representativeness. Furthermore, because the typical longitudinal panel maintains a static sample over time, sample size is limited, preventing time aggregation to increase sample representativeness over longer time intervals. In contrast, repeated cross-sections can be aggregated over time, enabling researchers to study small sub-populations over coarser time intervals. As a result, repeated cross-sections are far more common than longitudinal panels in tracking studies (Hsiao, 2007; Tourangeau, 2003).

In light of the above, our focus in this paper is *not* on tracking studies whose main interest is in estimating individual-level dynamics (which

* Corresponding author at: Melcher Hall 375E, 4800 Calhoun Rd., Bauer College of Business, University of Houston, Houston, TX 77204-6021, United States. Tel.: +1 713 743 9277; fax: +1 713 743 4572.
E-mail addresses: rexdu@bauer.uh.edu (R.Y. Du), kamakura@rice.edu (W.A. Kamakura).
[1] Tel.: +1 713 348 6307.

http://dx.doi.org/10.1016/j.ijresmar.2014.10.002
0167-8116/

require longitudinal panels). Rather, we focus on tracking studies whose main interest is in monitoring the 'state of population'. In particular, we focus on repeated cross-sectional surveys whose main goal is to monitor population means for variables of interest that are measured on interval or ratio scales. Such tracking studies are prevalent in marketing and virtually all the other social sciences, e.g., the Consumer Expenditure Survey and American Time Use Survey conducted by the U.S. Bureau of Labor Statistics, the Survey of Consumers conducted by Thomson Reuters-University of Michigan, and numerous syndicated trackers on product consumption, brand health and customer satisfaction that rely on repeated cross-sections drawn from omnibus panels maintained by large marketing research companies (e.g., Kantar Worldpanel, NPD, YouGov, Vision Critical).

The most common approach in dealing with repeated cross-sectional survey data is to pool the responses from all those interviewed in the same period, calculate the sample averages, and use those sample averages as estimates of population means in the corresponding time period. Such an approach is easy to implement but faces a major challenge — random sampling errors are confounded with genuine changes in population means. When one observes two sample averages from two different time periods, one does not know the extent to which the difference between these two sample averages is caused by sample composition differences or changes in population means. The former is purely a function of who were drawn into each sample and is therefore of little interest to the researcher. The latter is what the researcher is truly interested in uncovering.

The above confound is exacerbated when the sample size in each time period is small and the population is heterogeneous, as both factors lead to large random sampling errors. As sampling errors increase, true signals about population means become harder to detect, leading to not only more inaccuracies but also more volatility in the estimates of population means. In this paper, we develop a method that can substantially improve the statistical performance of tracking studies based on repeated cross-sections, by better separating random sampling errors from genuine changes in population means. To accomplish such a goal, we take full advantage of three patterns that are commonly embedded in repeated cross-sectional survey data:

1. Inter-temporal dependence in population means. While individual responses from non-overlapping cross-sections are independent over time, they reflect, collectively, the state of the population in each time period, which is obviously temporally dependent. For example, it should be rare that a brand's health would vary dramatically from one month to the next, even though the perceptions of individuals sampled in one month are independent from the perceptions of individuals sampled in the next month. By formally taking into account inter-temporal dependence in population states, our proposed model borrows information from all time periods in inferring the population means in any given period. This implies that, in estimating population means over time, our model smoothens the raw sample averages by filtering out larger-than-expected fluctuations (with the expected level of smoothness empirically determined), attributing these unusual shifts more to random sampling errors than to changes in population states.
2. Temporal co-movements among population means of multiple measures. In most tracking surveys, researchers gather data on multiple variables, many of which related to the same underlying constructs (e.g., customer attitudes with respect to different aspects of a product). To the extent that population means of these measures are manifestations of the same underlying population state and sample averages are manifestations of population means, the movements of the sample averages should be correlated over time. In other words, by formally taking into account temporal co-variations among sample averages, our proposed model borrows information across the sample averages of all measures in inferring the population means of any given measure. Intuitively, this implies that our model filters out idiosyncratic movements in any given measure's sample averages by triangulating them against how the other measures' sample averages move, with the expected pattern of co-movements empirically determined.
3. Cross-sectional co-variations across multiple measures. Due to factors such as common method bias, halo effect, heterogeneity in scale usage and other respondent characteristics, a respondent's answers to multiple questions from the same survey can be correlated with one another. When respondent-level data are available, such within-sample between-measure correlation will manifest itself and therefore can be uncovered from cross-sectional co-variations. However, if only sample averages were available, due to random differences in sample composition from one time period to another, cross-sectional between-measure co-variations would lead to spurious temporal co-movements among the measures' sample averages, which, unfortunately, would be confounded with genuine temporal co-movements in the measures' population means. In other words, in order to disentangle cross-sectional between-measure co-variations from temporal between-measure co-movements in population means, one needs to take advantage of tracking data at the respondent level, which our proposed model allows us to accomplish.

In the rest of the paper, we proceed as follows. We first review existing methods that can potentially be utilized to alleviate the confound between random sampling errors and genuine changes in population means, highlighting how each method leverages one or two of the three data patterns mentioned above. We then present our proposed approach, which, by simultaneously leveraging the three aforementioned patterns that are commonly embedded in repeated cross-sectional survey data, goes beyond all that has been attempted in the literature. To test our methodology and better understand the incremental value of leveraging each of the three data patterns, we conduct a Monte Carlo simulation using data from a known population from which we draw repeated cross-sections. Given that we know the true population means in each time period, we can make equitable comparisons in statistical performance across different approaches. After we thoroughly test the performance of our proposed model against known population means and benchmark models, we illustrate its use by applying it to data gathered through repeated cross-sectional surveys of customer perceptions and satisfaction for an automotive dealer, a situation that is often encountered by marketing researchers.

2. Tracking population dynamics with repeated cross-sectional samples

A seminal study on the analysis of repeated surveys was published by Scott and Smith (1974), who were the first to realize that while the observations from each wave of surveys might be independent, the population means being estimated from the sample averages could in fact be temporally dependent. Depending on the assumed inter-temporal dependency of the population means and the type of repeated cross-sectional sampling (overlapping or not), Scott and Smith (1974) and Scott, Smith, and Jones (1977) suggested different ARIMA models to better infer, from sample averages over time, the trend line of the population mean of a single response variable. They referred to this model-based approach for inferring population means over time as a *secondary analysis* of repeated cross-sectional data, as it relies on sample averages as inputs, as opposed to a *primary analysis* of raw respondent-level data. One recent example in the marketing literature of applying time series models to sample averages is Srinivasan, Vanhuele, and Pauwels (2010), who used a vector-autoregressive (VARX) market-response model in investigating the dynamics between marketing mix, brand sales and consumer mindset metrics, which were gathered through repeated cross-sectional surveys. Their focus, however, is not on better inferring population means from sample averages over time, and