JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY

JORS

John Boylan
Martin H. Kunc
Said Salhi
Zhe George Zhang

Journal of the Operational Research Society

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

Subject-to-group statistical comparison for open banking-type data

A. Svetlošák, M. de Carvalho & R. Calabrese

To cite this article: A. Svetlošák, M. de Carvalho & R. Calabrese (2023) Subject-to-group statistical comparison for open banking-type data, Journal of the Operational Research Society, 74:3, 703-718, DOI: 10.1080/01605682.2021.1952115

To link to this article: https://doi.org/10.1080/01605682.2021.1952115

Published online: 23 Aug 2021.

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
2023, VOL. 74, NO. 3, 703–718
https://doi.org/10.1080/01605682.2021.1952115

THE OPERATIONAL RESEARCH SOCIETY
Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

OPEN ACCESS
Check for updates

Subject-to-group statistical comparison for open banking-type data

A. Svetlošák[a,b], M. de Carvalho[b] and R. Calabrese[a]

[a]Business School of The University of Edinburgh, Edinburgh, UK; [b]School of Mathematics of the University of Edinburgh, Edinburgh, UK

ABSTRACT
Open banking (OB) creates an opportunity for financial institutions to offer more personalised services by better differentiating between a specific customer (*reference subject*) and similar customers (*comparison group*). We propose the time-varying comparative mean value as a statistical method that learns about the dynamics governing how the response of a reference subject differs from that of a comparison group, defined via covariate truncation. The proposed model can be regarded as a time-varying truncated covariate regression model of which a smooth version is devised by resorting to local polynomial regression. The simulation study suggests that our estimators accurately recover the true time-varying comparative mean value in a variety of scenarios. We showcase our methods using OB-type data from a financial service provider in the UK, with the dataset containing detailed information on customers' accounts across 70 UK financial institutions. By contrasting a specific customer against similar customers, our method offers interesting diagnostics that can be used by financial institutions to recommend personalised services.

ARTICLE HISTORY
Received 1 April 2020
Accepted 24 June 2021

KEYWORDS
F-barycenter; financial advice; nonparametric regression; open banking; personal finances; statistical planning; truncated regression

1. Introduction

In this paper we propose an approach for comparing a specific reference subject (e.g., a customer of a financial institution or an user of a financial application) to a comparison group. Since the proposed methodology is motivated from the recent open banking (OB) paradigm, we start by offering some background on the applied context motivating the methodological contributions.

The digital revolution of OB is impacting a wealth of decision-makers, problem-solvers, and innovators. In particular, under OB financial service providers can access datasets containing detailed financial information on potential customers across several financial institutions. Thanks to OB-type data financial institutions will have the opportunity to offer more personalised services by better differentiating between a specific customer (*reference subject*) and similar customers (*comparison group*). Trendsetters like Facebook (Cheng & Zagat, 2019) are already seising upon the opportunities stemming from subject-to-group comparison in an OB context, that is, differentiating between a reference subject and similar customers. Yet, thus far no methodological developments have been made that allow for a sound subject-to-group comparison, and one the goals of this paper will be to explore this open problem.

Different fields spreading from computer science (Wang et al., 2020), regulatory technologies (Buckley et al., 2020), and healthcare (Stranieri et al., 2021) are showing interest in OB. The digital revolution of OB is an initiative led by multiple governments, such as Australia, USA, UK, and EU in the form of the PSD2 legislation (Brodsky & Oakes, 2017; European Commission, 2018; He et al., 2020). The introduction of OB goes hand in hand with the currently expanding area of fintech, which aims to enhance customer experience (Gomber et al., 2018); this has led to a vast increase of personalised customer-centric fintech services (Breidbach et al., 2019). Fintech and OB are viewed as game-changing and disruptive innovations (Lee & Shin, 2018; Nicoletti et al., 2017; Omarini, 2018), bringing opportunities and challenges for financial institutions, consumers, academics, regulators, and governments alike.

Besides industry demand, the developments in OB and fintech have been accelerated by innovations in machine learning, artificial intelligence, and related fields (Mention, 2019). Yet, little has been done to address these new challenges from a statistical perspective. We aim to fill this methodological and applicational gap by providing a statistical solution to subject-to-group comparison, that provides

CONTACT A. Svetlošák andrejsvetlosak@gmail.com Business School of The University of Edinburgh, 29 Buccleuch Pl, Edinburgh EH8 9JS, UK

valuable OB-based financial advice that is intuitive and easy to interpret.

OB-type data has the potential to change how personalised financial services are provided, and creates a necessity to develop subject-to-group comparison methods tailored for these purposes. Unlike subject-to-group comparisons, methods for comparing population groups, such as hypothesis tests and regressions with group indicators, are well-studied. Recent micro-financial studies utilise new and highly personalised data sources to provide valuable insight on credit scoring (De Cnudde et al., 2019; Óskarsdóttir et al., 2019; Zhu et al., 2020). Yet these methods are not suitable for identifying and analysing the differences between a specific customer and a group with similar characteristics, and thus can not harness the full potential of OB-type data and alike. A key goal of this paper is to propose an approach that can be used for learning about the latter differences.

In this paper, we contribute by providing an approach for comparing a specific reference subject to a comparison group defined via covariate truncation, based on what we will refer to as the *time-varying comparative mean value*. While the proposed methodology is motivated from OB data, we underscore that our framework is general, and that thus the proposed methods can be applied to any context where the goal is to conduct a subject-to-group comparison. To the best of our knowledge, our paper pioneers the mathematical modelling of subject-to-group comparisons.

Here and below we will use the expression covariate truncation to refer to a truncation centred on the reference subject's covariate value. Conditioned on the comparison group identified by covariate truncation (say, income) the expected value of the response (say, expenditure), is analysed. The homogeneity in terms of the covariate within the comparison group is controlled by a similarity variable (δ below) set by the user; we analyse the properties of the proposed method in terms of this variable. No parametric structure on the studied variables is assumed. A smooth time-varying comparative mean value is also devised by resorting to local polynomial regression (Fan, 1996). By contrasting the time-varying comparative mean value with the reference subject values, we can identify differences between the reference subject and a group with similar characteristics (peers). By varying the similarity variable (δ) information on the distribution of the response can be obtained.

A statistical concept that relates to the here proposed time-varying comparative mean value is the so-called F-barycenter (Hill & Monticino, 1998). For a random variable $X \sim F$, with finite expectation and with strictly increasing F, the F-barycenter of $(a, c]$ is defined as

$$b_F = \mathrm{E}\{X | X \in (a, c]\}. \tag{1}$$

As we will elaborate below, the here proposed time-varying comparative mean value is designed for a regression framework, and it can be regarded as an $F_{Y_t|X_t}$-barycenter, for the pairs of stochastic processes $\{X_t\}$ and $\{Y_t\}$, and with $(a, c]$ centred at the reference subject's covariate.

Other statistical methods that conceptually relate to the here proposed time-varying comparative mean value are the marginal expected shortfall (MES) and the systemic expected shortfall (SES) proposed by Acharya et al. (2010) and further developed in Cai et al. (2015) and Acharya et al. (2017); we will provide further details on these methods in Section 2.4. Although the related statistical concepts have some points of contact with our methods, their approach would not be suitable for offering a subject-to-group comparison. The time-varying comparative mean value also relates to regression towards the mean and k-nearest neighbours. Yet the latter two methods converge in probability to a conditional mean given a fixed value of the covariate (Hastie et al., 2001, Section 2.4), whereas here the goal will be on learning about a conditional mean—given the covariate is in an interval induced by covariate truncation. The time-varying comparative mean value is also connected to truncated regression methods, such as those developed in Efron and Petrosian (1999), Shen (2010), and Ying et al. (2019). We show that the time-varying comparative mean value can be seen as a regression with truncated covariates. This is an important distinction to the latter regression methods, as those methods are tailored for truncated responses. To our knowledge, the comparative mean value provides a first attempt to model truncated covariates on a regression framework.

The remainder of this paper unfolds as follows. In Section 2 we present the comparative mean value (static and time-varying versions), introduce estimation methods, show the connection of the comparative mean value to regression, and provide comments on related statistical concepts. In Section 3 we test the estimators of the time-varying comparative mean value using a Monte Carlo simulation. In Section 4 we then showcase the method's use on Money Dashboard personal financial data of UK residents in 2017 and 2018. We track the development of expenditure of selected reference subjects, contrasted with the expenses of peers with similar income. In Section 5 we summarise and discuss our main findings.

2. Comparative mean value

2.1. Comparative mean value: Construction and properties

Let $X \sim F_X$ and $Y \sim F_Y$, with F_X and F_Y denoting the distribution functions of X and Y, respectively; let

$X_0 = x_0$ and $Y_0 = y_0$ be fixed values for a reference subject. Although below we define the comparative mean value as a general concept, for our applied setup the expression "reference subject" should be understood as representing a customer of a financial institution or an user of a financial application. We are interested in the expected value of the response Y, for subjects that are similar to the reference subject in respect to the covariate X. Specifically, the goal is contrasting the reference subject to similar subjects, whose covariate X lies within a neighbourhood of x_0, $W_\delta(x_0) \equiv [x_0 - \delta, x_0 + \delta]$. All comparisons below are for $\delta > 0$ such that for all $x \in W_\delta(x_0)$ it holds that $f_X(x) = \mathrm{d}F_X/\mathrm{d}x > 0$; the extension to the asymmetric setting, $W_{\delta_1,\delta_2}(x_0) = [x_0 - \delta_1, x_0 + \delta_2]$, is straightforward but notationally burdensome. The modelled object of interest is what we refer to as the *comparative mean value* (at level δ)

$$\mu_\delta = \mathrm{E}\{Y | X \in W_\delta(x_0)\}. \tag{2}$$

As $\delta \to 0$, there are connections with regression towards the mean, and we explore these links in Section 2.3.

The comparative mean value (2) obeys similar properties as the F-barycenter (Hill & Monticino, 1998, Lemma 2.2); indeed it follows that:

$$\begin{aligned}\{F_X(x_0 + \delta) - F_X(x_0 - \delta)\}\mu_\delta \\ = \{F_X(\mu_\delta) - F_X(x_0 - \delta)\}\mathrm{E}\{Y | X \in [x_0 - \delta, \mu_\delta]\} \\ + \{F_X(x_0 + \delta) - F_X(\mu_\delta)\}\mathrm{E}\{Y | X \in [\mu_\delta, x_0 + \delta]\},\end{aligned} \tag{3}$$

and $\mathrm{E}\{Y | X \in [x_0 - \delta, \mu_\delta]\} = \mu_\delta$ if and only if $\mathrm{E}\{Y | X \in [\mu_\delta, x_0 + \delta]\} = \mu_\delta$.

In practice, we estimate μ_δ as follows. Let $n+1$ be the total number of subjects, for which data $\{X_i, Y_i\}_{i=0}^n$ are available. We define the *window of comparison* as the set $A_\delta = \{i : X_i \in W_\delta(x_0)\}$. This means that we are conditioning on those subjects whose covariates X_i are within a range of δ from the reference value x_0.

Let $k_{n,\delta}$ be the cardinality of the window of comparison, that is $k_{n,\delta} = |A_\delta|$; although k is a function of both n and δ, to ease notation we omit these dependencies and write $k \equiv k_{n,\delta}$. We define the *comparative sample mean* (at level δ) as

$$\hat{\mu}_\delta = \frac{1}{k}\sum_{i \in A_\delta} Y_i. \tag{4}$$

We can see that the estimator in (4) first selects subjects whose covariates X_i are within the interval $W_\delta(x_0)$, centred around the reference subject's covariate x_0; then the average of the response Y_i for the selected peer group is calculated. Assume:

$$1. \lim_{\delta \to 0} k = |A_0| \quad 2. \lim_{\delta \to \infty} k = n + 1 \quad 3. \lim_{n \to \infty} k = \infty.$$

Under the latter assumptions it can be shown that (4) has the following properties

$$1. \lim_{\delta \to 0} \hat{\mu}_\delta = y_0, \text{a.s.} \quad 2. \lim_{\delta \to \infty} \hat{\mu}_\delta = \bar{Y}, \text{a.s.} \quad 3. \hat{\mu}_\delta \underset{p}{\rightarrow} \mu_\delta \tag{5}$$

as $n \to \infty$, for $\delta > 0$. Next, we devise a time-varying framework extending the methods introduced above.

2.2. Time-varying comparative mean value

To track the dynamics governing the comparative mean value, we now develop a time-varying version. Let $\{X_t\}$ and $\{Y_t\}$ be stochastic processes and let $\{X_{0,t} = x_{0,t}\}$ and $\{Y_{0,t} = y_{0,t}\}$ be the fixed values of the reference subject. Our object of interest is now the *time-varying comparative mean value* (at level δ)

$$\mu_{\delta,t} = \mathrm{E}\{Y_t | X_t \in W_\delta(x_{0,t})\}. \tag{6}$$

The time-varying version of properties (3) hold for (6). Let $n_t + 1$ be the total number of subjects in period t and let

$$\{X_{i,1}, Y_{i,1}\}_{i=0}^{n_1}, \ldots, \{X_{i,T}, Y_{i,T}\}_{i=0}^{n_T}$$

be the observed data (say, income and expenditure). We again fix the reference subject's index at $i = 0$. The *time-varying window of comparison* is $A_{\delta,t} = \{i : X_{i,t} \in W_\delta(x_{0,t})\}$. Denoting $k_t = |A_{\delta,t}|$, we assume

$$1. \lim_{\delta \to 0} k_t = |A_{0,t}| \quad 2. \lim_{\delta \to \infty} k_t = n_t + 1$$
$$3. \lim_{n_t \to \infty} k_t = \infty.$$

The estimator of (6) is the *time-varying comparative sample mean* (at level δ)

$$\hat{\mu}_{\delta,t} = \frac{1}{k_t}\sum_{i \in A_{\delta,t}} Y_{i,t}. \tag{7}$$

For any time period, subjects whose covariates $X_{i,t}$ are within the interval $W_\delta(x_{0,t})$ are selected and the conditional mean of the response Y_t is calculated. Let $\bar{Y}_t = 1/n_t \sum_{i=1}^{n_t} Y_{i,t}$. The estimator $\hat{\mu}_{\delta,t}$ has the following properties

$$\begin{aligned}&1. \lim_{\delta \to 0} \hat{\mu}_{\delta,t} = y_{0,t}, \text{a.s.} \quad 2. \lim_{\delta \to \infty} \hat{\mu}_{\delta,t} = \bar{Y}_t, \text{a.s.} \\ &3. \hat{\mu}_{\delta,t} \underset{p}{\rightarrow} \mu_{\delta,t}\end{aligned} \tag{8}$$

as $n_t \to \infty$, for $\delta > 0$. We now derive a smooth version of the time-varying comparative sample mean so to smooth the dynamics. For this purpose we resort to local polynomial regression (Fan, 1996) as the obtained smooth dynamics naturally extend those that would be obtained via standard regression methods, in the sense that they can be understood as a running weighted regression. When the degree of the polynomial is zero, we get the well-known Nadaraya–Watson estimator (Watson, 1964). Using a nonparametric approach provides two main benefits. First, we obtain a predictor for the time-varying comparative mean value. Second, while subject-to-group comparison contains valuable information, if the empirical results oscillate too much in time it is hard to recognise a clear trend. Smoothing the results makes them more intuitive and easy to understand.

The nonparametric regression model is defined as $\mu_{\delta,t} = m_t(\delta) + \varepsilon_t$, where ε_t is an independent and identically distributed error term with $\mathrm{E}\{\varepsilon_t\} = 0$ and $\mathrm{var}\{\varepsilon_t\} = \sigma^2$. The function $m_t(\delta)$ can be estimated by what we refer to as the *smooth time-varying comparative sample mean* (at level δ)

$$\widehat{m_t}(\delta) = \frac{\sum_{i=1}^T K_h(t-i)\hat{\mu}_{\delta,i}}{\sum_{i=1}^T K_h(t-i)}. \tag{9}$$

In (9), $K_h(\cdot) = K(\cdot\ /h)\ /h$, where K is a kernel function and $h > 0$ is a smoothing parameter (bandwidth). As usual, the bandwidth $h = h_T$ is a sequence such that $h \to 0$ and $hT \to \infty$, as $T \to \infty$. It is well known that the choice of the kernel has little effect on the estimates (Wand, 1995, Section 2.7). Yet the choice of bandwidth is an important one, as a poor choice can lead to over-smoothing or under-smoothing. A standard approach to determine h is via cross-validation, with least-squares based validation being regarded as optimal (DasGupta, 2008, Section 3.10.2). We use the method proposed by Li and Racine (2004) and implemented by the authors in the R package np. Under fixed n_t and regularity conditions, the estimator in (9) inherits the following asymptotic properties from the Nadaraya–Watson estimator (e.g., Racine, 2001, Section 2), as $T \to \infty$

$$\mathrm{bias}\{\widehat{m_t}(\delta)\} \approx m_t^{(2)}(\delta)\frac{h^2}{2}\int_{-\infty}^{\infty} K(v)v^2\ \mathrm{d}v,$$

$$\mathrm{var}\{\widehat{m_t}(\delta)\} \approx \frac{\sigma^2}{Th}||K||_2^2.$$

Here $||\cdot||_2$ is the L^2 norm; noticeable in the latter expressions is the trade-off between variance and bias in respect to the bandwidth.

2.3. Time-varying comparative mean value as a truncated covariate regression

We now discuss the connection between the time-varying comparative mean value (6) and regression, which originates from observing that

$$\mu_{\delta,t} \to \mathrm{E}\{Y_t|X_t = x_{0,t}\}, \qquad \text{as } \delta \to 0.$$

For generality, we first focus on the nonparametric specification $Y_t = m_t(X_t) + \varepsilon_t$. For any $\delta \geq 0$, we then have

$$\mu_{\delta,t} = \mathrm{E}\{m_t(X_t)|X_t \in W_\delta(x_{0,t})\} + \mathrm{E}\{\varepsilon_t|X_t \in W_\delta(x_{0,t})\}.$$

Denoting by F_t the cumulative distribution function of the covariate X_t, and using Ruiz and Navarro (1996, p. 564), we can provide a general formula for computing the time-varying comparative mean value, only assuming that X_t and ε_t are independent and $\mathrm{E}(\varepsilon_t|X_t) = 0$. Under these assumptions it can be shown that $\mathrm{E}\{\varepsilon_t|X_t \in W_\delta(x_{0,t})\} = 0$. Thus, for $\delta > 0$ the time-varying comparative mean value (6) is

$$\mu_{\delta,t} = \frac{1}{F_t(x_{0,t}+\delta) - F_t(x_{0,t}-\delta)}\int_{x_{0,t}-\delta}^{x_{0,t}+\delta} m_t(x)\ \mathrm{d}F_t(x). \tag{10}$$

Equation (10) suggests another route for estimating the time-varying comparative mean value. Yet it would require estimation of the margins and of the conditional mean, followed by integration, and thus it is not pursued further here.

If the association between the response Y_t and the covariate X_t is linear, that is if $Y_t = \alpha_t + \beta_t X_t + \varepsilon_t$, the time-varying comparative mean value (6) becomes

$$\begin{aligned}\mu_{\delta,t} &= \alpha_t + \beta_t \mathrm{E}\{X_t|X_t \in W_\delta(x_{0,t})\} + \mathrm{E}\{\varepsilon_t|X_t \in W_\delta(x_{0,t})\}\\ &= \alpha_t + \beta_t\frac{1}{F_t(x_{0,t}+\delta) - F_t(x_{0,t}-\delta)}\int_{x_{0,t}-\delta}^{x_{0,t}+\delta} x\ \mathrm{d}F_t(x).\end{aligned} \tag{11}$$

For specific parametric distributions we can provide closed form expressions for $\mu_{t,\delta}$. In Table 1, adapted from Johnson et al. (2005, p. 134) and Ruiz and Navarro (1996, p. 570), we do this for a selection of distributions, under the linear specification $Y_t = \alpha_t + \beta_t X_t + \varepsilon_t$. Noteworthy is the case $X_t \sim \mathrm{Unif}[a,b]$, for which the comparative mean value (11) is independent of δ. We depict a linear setting with $X_{i,t} \sim N(\theta_t, \sigma_t)$ in Section 3.1, Scenario B.

2.4. Comments on conceptually-related methods

Beyond the links with the F-barycenter (1) highlighted in the Introduction, we now make some remarks on conceptually-related methods. As discussed in the Introduction, the other concepts relating to the comparative mean value are the MES and the SES. Time-varying version of all these concepts can be easily defined; yet to streamline the discussion we will focus on the static setup, and thus we work with pairs of random variables (X, Y) rather than stochastic processes (X_t, Y_t). The MES is defined as

$$\mathrm{MES}(p) = \mathrm{E}\{Y|X > Q(1-p)\}, \quad 0 < p < 1,$$

where $Q(p) = \inf\{x : F(x) \geq p\}$; the SES has links with the MES and details on it can be found in Acharya et al. (2017). While keeping in mind the

Table 1. Parametric examples of the time-varying comparative mean value for linear dependence between Y_t and X_t.

Distribution	Time-varying comparative mean value	Restrictions
$N(\theta_t, \sigma_t^2)$	$\alpha_t + \beta_t\{\theta_t + \sigma_t \frac{\phi\left(\frac{x_{0,t}-\delta-\theta_t}{\sigma_t}\right) - \phi\left(\frac{x_{0,t}+\delta-\theta_t}{\sigma_t}\right)}{\Phi\left(\frac{x_{0,t}+\delta-\theta_t}{\sigma_t}\right) - \Phi\left(\frac{x_{0,t}-\delta-\theta_t}{\sigma_t}\right)}\}$	$\delta \neq 0$
Unif(a, b)	$\alpha_t + \beta_t x_{0,t}$	$a<b$ and $\delta \neq 0$
Exp (λ_t)	$\alpha_t + \beta_t \frac{exp\,(2\lambda_t\delta)(x_{0,t}+\delta+1/\lambda_t)-x_{0,t}-\delta-1/\lambda_t}{exp\,(2\lambda_t\delta)-1}$	$\lambda_t>0$ *and* $\delta \neq 0$
Power(c_t)	$\alpha_t + \frac{\beta_t}{c_t+1} \frac{\{c_t(x_{0,t}+\delta)+a\}(x_{0,t}+\delta-a)^{c_t} - \{c_t(x_{0,t}-\delta)+a\}(x_{0,t}-\delta-a)^{c_t}}{(x_{0,t}+\delta-a)^{c_t} - (x_{0,t}-\delta-a)^{c_t}}$ $a \leq x_{0,t}-\delta<x_{0,t}+\delta \leq b$ *and* $c_t>0$	$a \leq x_{0,t}-\delta < x_{0,t}+\delta \leq b$ *and* $c_t > 0$
Cauchy (location $= \theta_t$, scale $= \sigma_t$)	$\alpha_t + \beta_t\theta_t + \frac{\beta_t\sigma_t}{2} \frac{log\,(\sigma_t^2+(x_{0,t}+\delta-\theta_t)^2) - log\,(\sigma_t^2+(x_{0,t}-\delta-\theta_t)^2)}{\arctan(\frac{x_{0,t}+\delta-\theta_t}{\sigma_t}) - \arctan(\frac{x_{0,t}-\delta-\theta_t}{\sigma_t})}$	$\sigma_t > 0$ *and* $\delta \neq 0$
Pareto(scale $= \sigma_t$, shape $= \theta_t$)	$\alpha_t + \beta_t\{$ $\frac{\theta_t}{1-\theta_t} \frac{(x_{0,t}-\delta)^{\theta_t}(x_{0,t}+\delta) - (x_{0,t}+\delta)^{\theta_t}(x_{0,t}-\delta)}{(x_{0,t}+\delta)^{\theta_t} - (x_{0,t}-\delta)^{\theta_t}}$ $\frac{x_{0,t}^2-\delta^2}{\delta^2} log\,(\frac{x_{0,t}+\delta}{x_{0,t}-\delta})$	$\theta_t \neq 1$ *and* $\delta \neq 0$ $\delta \neq 0$ *and* $x_{0,t}\pm\delta \neq 0$
Laplace (location $= \theta_t$, scale $= \sigma_t$)	$\frac{exp\,(\frac{2\delta}{\sigma_t})(\sigma_t + x_{0,t}-\delta)-\sigma_t-x_{0,t}-\delta}{exp\,(\frac{2\delta}{\sigma_t})-1}$ $\alpha_t + \beta_t\{ \frac{2\theta_t-(x_{0,t}-\delta-\sigma_t)\,exp\,(\frac{x_{0,t}-\delta-\theta_t}{\sigma_t})-(x_{0,t}+\delta+\sigma_t)\,exp\,(\frac{\theta_t-x_{0,t}-\delta}{\sigma_t})}{2-\,exp\,(-\frac{x_{0,t}+\delta-\theta_t}{\sigma_t}) - \,exp\,(\frac{x_{0,t}-\delta-\theta_t}{\sigma_t})}$ $\frac{exp\,(\frac{2\delta}{\sigma_t})(x_{0,t}+\delta-\sigma_t)-x_{0,t}+\delta+\sigma_t}{exp\,(\frac{2\delta}{\sigma_t})-1}$	$\sigma_t>0, \theta_t \leq x_{0,t}-\delta$ *and* $\delta \neq 0$ $\sigma_t>0$ *and* $x_{0,t}-\delta<\theta_t \leq x_{0,t}+\delta$ $\sigma_t>0, x_{0,t}+\delta<\theta_t$ *and* $\delta \neq 0$

different context in which these concepts are devised, namely the evaluation of risk for financial institutions, the main difference between the MES, the SES and the time-varying comparative mean value (6) is the set the expected value is conditioned on. The MES and SES are conditioned on unbounded half-open intervals. However, for our purpose of comparing a reference subject to a group of peers it is more meaningful to condition on a bounded interval around the reference subject covariate.

Ex-ante one could be led to believe that k-nearest neighbours would be tantamount to our time-varying comparative mean value. Yet as we discuss below, k-nearest neighbours is designed for estimating $E\{Y|X = x_0\}$—similarly to regression towards the mean methods—whereas our approach is tailored for estimating $E\{Y|X \in W_\delta(x_0)\}$. To see this, let $N_k(x_0)$ be the set of k closest values of X_i to x_0. For k-nearest neighbours it can be shown that (Hastie et al., 2001, Section 2.4)

$$\frac{1}{k}\sum_{X_i \in N_k(x_0)} Y_i \underset{p}{\rightarrow} \mathrm{E}\{Y|X = x_0\}, \quad \text{as } n, k \rightarrow \infty,$$

assuming $k/n \rightarrow 0$. Whereas, as noted in (5), for the comparative mean value we have

$$\frac{1}{k}\sum_{i \in A_\delta} Y_i \underset{p}{\rightarrow} \mathrm{E}\{Y|X \in W_\delta(x_0)\}, \quad \text{as } n \rightarrow \infty.$$

In words, the time-varying comparative mean value converges in probability to an expected value conditioned on the interval $W_\delta(x_0)$ induced by covariate truncation, while k-nearest neighbours converge in probability to an expected valued conditioned on a fixed value of the covariate. Additionally, as $n \rightarrow \infty$ we have $k \rightarrow \infty$, but observe that if $\delta \rightarrow \infty$, k is not an intermediate sequence as $k/n = 1$.

Another area of research related to the time-varying comparative mean value is truncated regression. In certain applications it is necessary to truncate the observations and focus only on a selected interval. Based on the results from, Turnbull (1976), Efron and Petrosian (1999) propose the nonparametric maximum likelihood estimate (NPMLE) for doubly truncated data. The authors study a truncated response Y, restricted to a closed interval $[a, c]$ and covariate X. Using the same setting, Shen (2010) provide an alternative derivation of the NPMLE hazard function, and recently Ying et al. (2019) provided estimators for the standard linear model. As shown in Section 2.3, the time-varying comparative mean value (6) can be thought as a regression, in which the covariate is truncated to the window of comparison. An important distinction is that the time-varying comparative mean value is for truncated covariates, whereas the latter regression methods are tailored for truncated responses. Indeed, to our knowledge the literature covering truncated regression has focussed on truncated responses, while the time-varying comparative mean value provides a first attempt to model truncated covariates on a regression framework.

Table 2. Data generating processes used over the simulation study; we use the shape-scale parametrisation of the Gamma distribution.

	Data generating process	
Data	Scenario A	Scenario B
Covariate $(X_{i,t})$	$\text{Gamma}(\exp(t/100), \exp(t/100))$	$N(\lvert\sin(t/10)+1\rvert, 5)$
Response $(Y_{i,t})$	$\text{Gamma}(\{\sin(t/10)+t/10\}^2/t, 2)$	$5+2\,X_{i,t}$

3. Simulation study

3.1. Monte Carlo simulation

In this section we assess the finite sample performance of the time-varying comparative sample mean, as defined in (7) and the smooth time-varying comparative sample mean, as defined in (9). Before reporting the results of the Monte Carlo simulation, we first illustrate the methods on a single run experiment. For this purpose we generate data $\{X_{i,1}, Y_{i,1}\}_{i=0}^{n_1}, ..., \{X_{i,T}, Y_{i,T}\}_{i=0}^{n_T}$ in two scenarios and for two values of $n_t = n$. In Scenario A, we generate independent $X_{i,t}$ and $Y_{i,t}$. In Scenario B, we generate $X_{i,t}$ and the $Y_{i,t}$ are obtained by the following linear dependency structure $Y_{i,t} = \alpha + \beta X_{i,t}$. The parametrisation of both scenarios is available from Table 2.

In both scenarios we set the reference subject's values to be $x_{0,t} = \text{E}(X_t)$ and $y_{0,t} = \text{E}\{Y_t|X_t \in W_\delta(x_{0,t})\}$. For Scenario B, the comparative mean value in this case is given by the parametric formulation of (11) for the normal distribution found in Table 1. Because of our choice of reference value $x_{0,t} = \text{E}(X_{i,t})$, and the symmetry of the normal distribution density function the time-varying comparative mean value (6) in Scenario B is $\mu_{\delta,t} = \text{E}\{Y_t|X_t \in W_{\delta,}(x_{0,t})\} = 5 + 2|\sin(t/10) + 1|$.

Both scenarios are tested for $n_t = 1,000$ and $n_t = 10,000$. In the Monte Carlo experiment we repeat the single run simulation 1, 000 times. Two additional scenarios with jumps in the response are presented in the Supplementary Material. As mentioned in Section 2.2 cross-validation by Li and Racine (2004) is used to select the smoothing bandwidth. However, to avoid overfitting the asymptotically optimal bandwidth for kernel density methods given by DasGupta (2008, p. 531) is set as a lower bound.

As expected, based on properties listed in (8), in Figures 1 and 2 the following dynamics can be observed. The number of subjects with $X_{i,t}$ in $A_{\delta,t}$ increases as either δ or n_t increase. For $\delta \to 0$, only subjects with $X_{i,t} = x_{0,t}$ are in the window of comparison $A_{\delta,t}$. For continuous variables X_t and Y_t this is the case only for the reference subject. Thus, in both scenarios we get $\hat{\mu}_{0,t} = y_{0,t}$. As we slightly increase δ, we get more subjects in the window of comparison $A_{\delta,t}$ and the time-varying comparative sample mean fluctuates around the $\mu_{\delta,t}$. As either δ or n_t increase we get that $\hat{\mu}_{\delta,t}$ approaches $\mu_{\delta,t}$. The single run experiment and the Monte Carlo simulation indicate that both the time-varying comparative sample mean (7) and the smooth time-varying comparative sample mean (9) model the time-varying comparative mean value (6) well. Furthermore, the estimates get closer to the real value as δ and/or n_t increase. This is due to more subjects being in the window of comparison, as can be seen on the x-axis in Figures 1 and 2.

3.2. Comparative mean value over all windows of comparison

We now examine how the time-varying comparative mean value looks like when we consider all windows of comparison. For this purpose we fix the time at $t = 50$ and plot the time-varying comparative sample mean (7) as a function of δ. The obtained results, as seen in Figure 3, provide information on the position of the reference subject's response among all comparison groups selected using the window of comparison, as well as the total sample mean.

From (8), we know that for $\delta \to 0$ we have $\hat{\mu}_{0,t} = y_{0,t}$. This means that the position of the time-varying comparative sample mean (7) as a function of δ, at $\delta = 0$, provides us information on the position of the reference subject compared to the total sample mean. The greater the absolute slope at small values of $\delta>0$, the more the reference subject's response deviates from the average in the comparison group. For large positive slopes the subject's response $y_{0,t}$ is considerably lower, than the responses $Y_{i,t}$ of close peers. For small negative slopes, the subject's response is considerably higher, than the response average of close peers. As δ increases more subjects are included in the window of comparison, and the comparative sample mean approaches the total sample mean.

To provide a specific example of interpreting the time-varying comparative mean value over all windows of comparison, we choose to focus on Scenario A. In Figure 3, Scenario A, we can see the reference subject's value is just below the sample mean. Per construction the reference subject's $y_{0,50}$ gives us the time-varying comparative mean value (6) at $t = 50$. The difference between the time-varying comparative mean value and the total sample

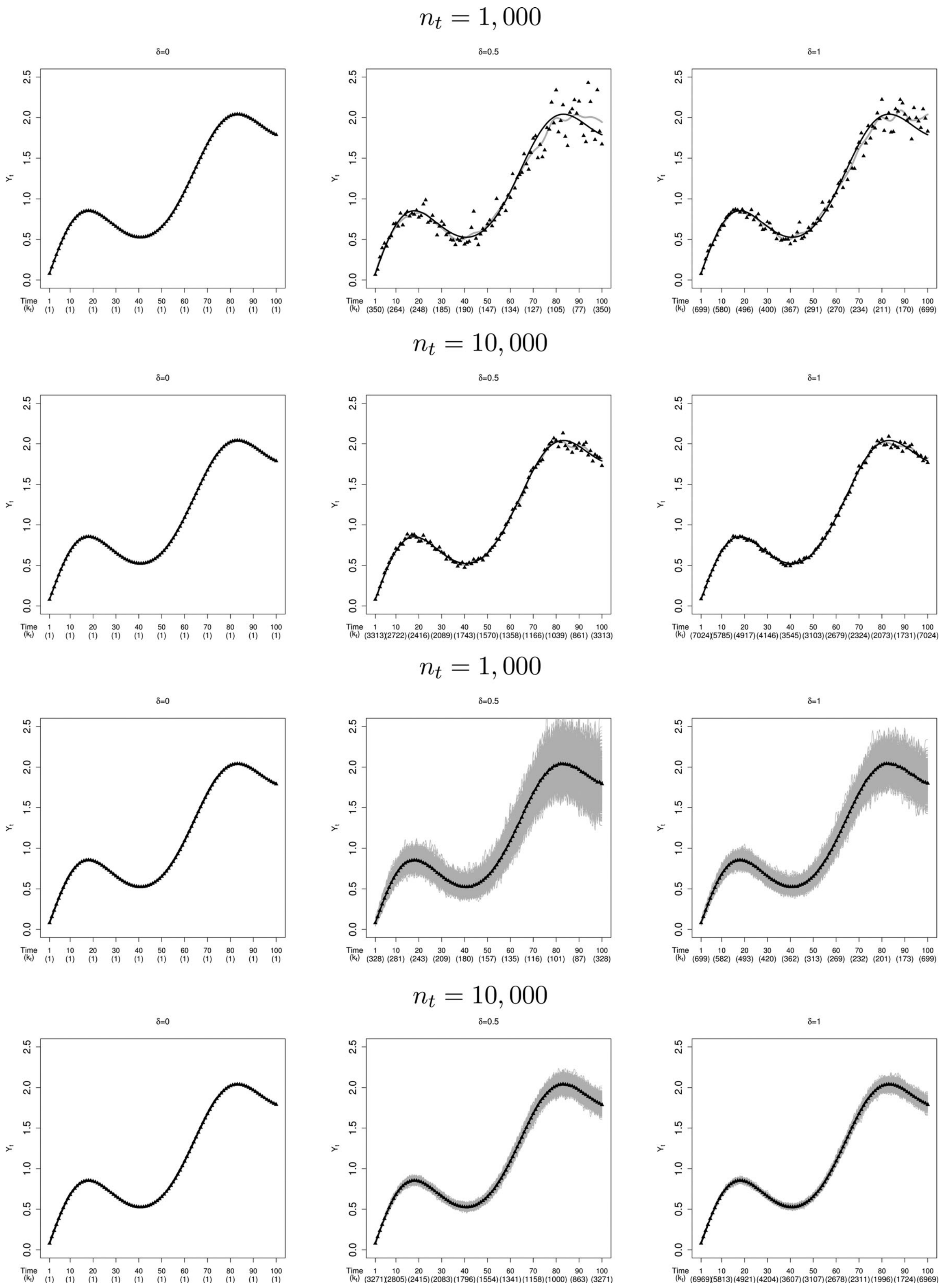

Figure 1. Time-varying comparative mean value, Scenario A: Top two rows show the single run experiment, and the bottom two rows the Monte Carlo simulation. The time-varying comparative sample mean (7) is represented by the triangles (▲). The grey curves represent the customer smooth time-varying comparative sample means (9), and the black curve is the true comparative mean value (6). The *x*-axis denotes time, and the number of subjects in the window of comparison, k_t.

mean is caused by the variance in each scenario, which is particularly noticeable in Scenario B where $\text{var}(Y_{i,t}) = 100$, and disappears at sufficiently high sample sizes. As δ slightly increases from the starting point $\delta = 0$ the time-varying comparative sample mean jumps up and down. This means, there are subjects with very similar $X_{i,50}$, but very different $Y_{i,50}$ values, compared to the

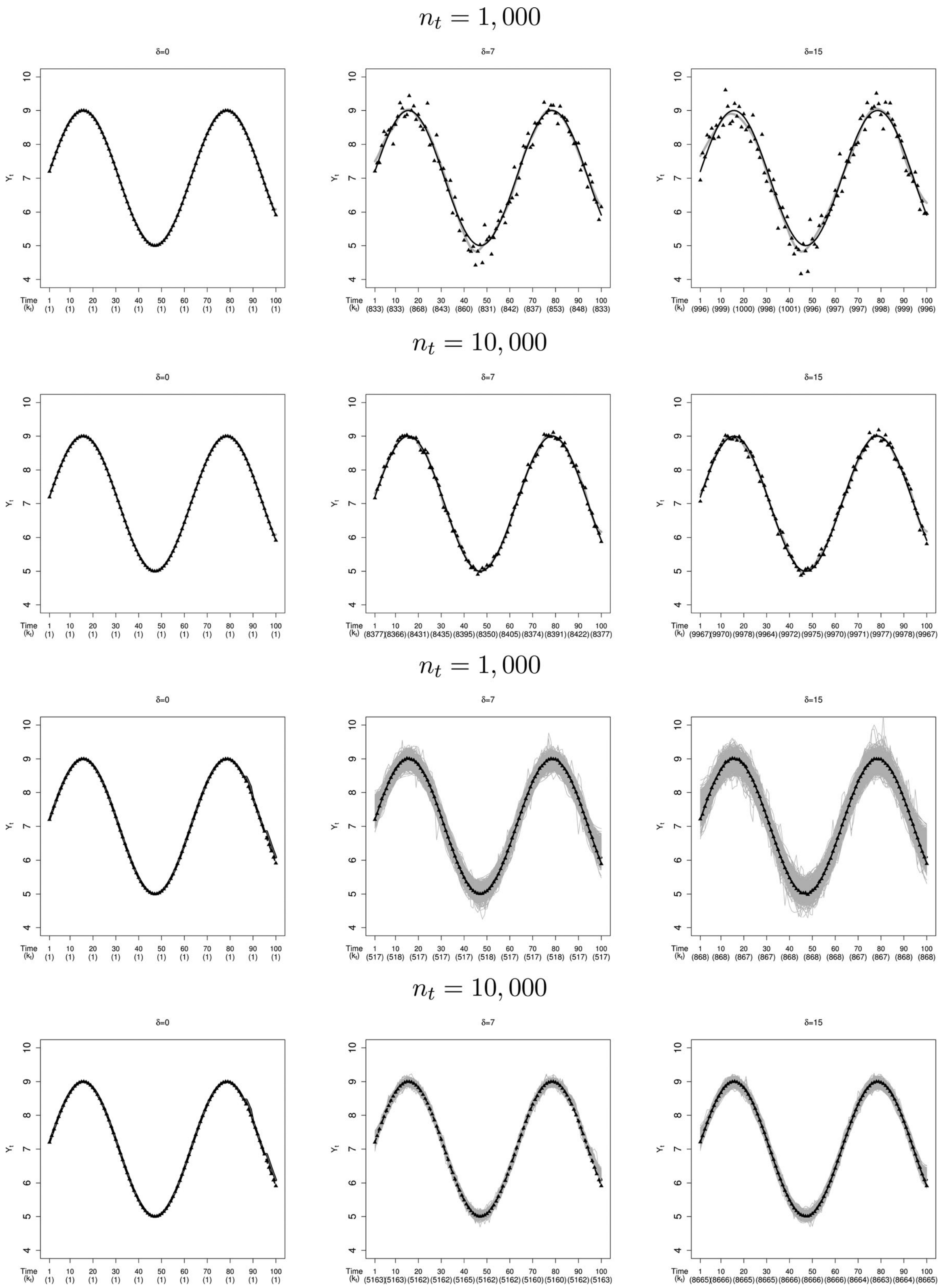

Figure 2. Time-varying comparative mean value, Scenario B: Top two rows show the single run experiment, and the bottom two rows the Monte Carlo simulation. The time-varying comparative sample mean (7) is represented by the triangles (▲). The grey curves represent the customer smooth time-varying comparative sample means (9), and the black curve is the true comparative mean value (6). The *x*-axis denotes time, and the number of subjects in the window of comparison, k_t.

response of the reference subject. The intensity of these fluctuations at small values of δ is due to relatively low number of peers in the window of comparison and the robustness properties of the mean. For δ around $(0.1, 2.2)$ the time-varying comparative sample mean stays below the reference subject's $y_{0,50}$; this indicates that the reference subject's response is higher than the average in any other group of peers created by the window of comparison $A_{\delta,50}$ for δ around $(0.1, 2.2)$. Such a subject could be considered to be an outlier in the latter peer groups. After $\delta>2.2$ the time-varying

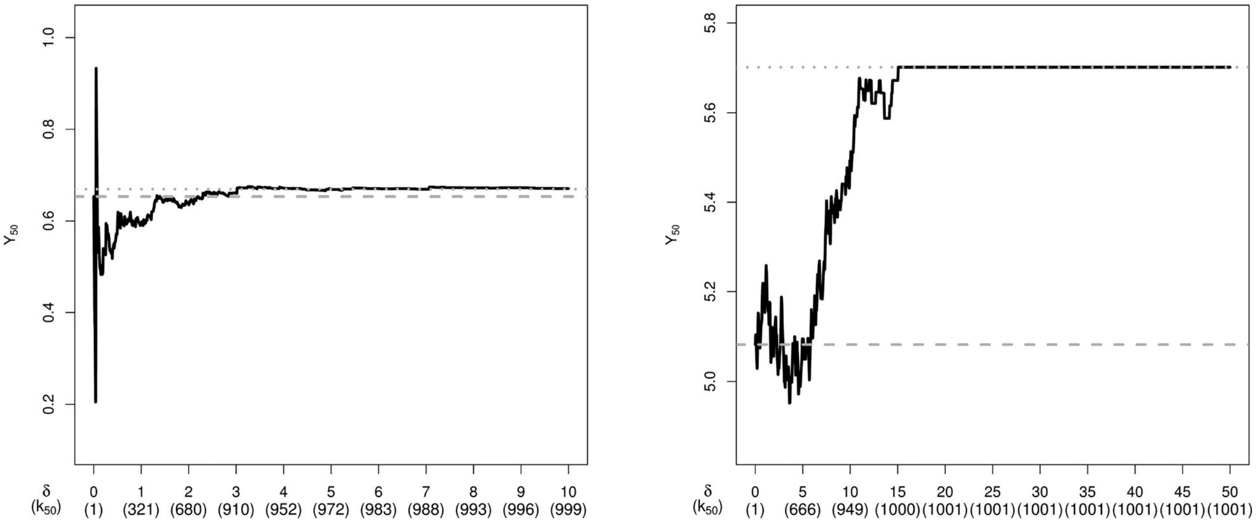

Figure 3. Time-varying comparative mean value over all windows of comparison for Scenario A (left) and Scenario B (right): The black curve represents the time-varying comparative sample mean (7) as a function of δ, and the dotted line is the total sample mean and the dashed line is the time-varying comparative mean value (6) for $t = 50$. The x-axis shows the δ values and the number of subjects in the window of comparison, k_t.

comparative sample mean as a function of δ aligns with the total sample mean, which is just slightly higher than the time-varying comparative mean value, indicating subjects with similar $X_{i,50}$ values have on average slightly higher $Y_{i,50}$ values.

4. Empirical analysis

4.1. Data description

We showcase the usefulness of the comparative mean value by considering three reference subjects, and comparing their expenditure to subjects with similar income.

The concepts and methods devised here were motivated by a collaboration with a UK financial services provider, Money Dashboard (MD), and the implementation of the time-varying concept in financial analysis platforms and financial applications. This is naturally connected to the high demand for personalised financial advice in the UK. According to the Financial Conduct Authority (2017, p. 10), in the last 12 months 6% of UK adults have received financial advice on personal investments, with "45% of those who have not received advice in the last 12 months report that they have had regulated financial advice related to investments, saving into a pension or retirement planning in the past."

MD supplies its customers (in this Section users) with a summary of their registered accounts at any financial institution in the UK via an application. Further details on the business lines of MD can be found at www.moneydashboard.com

The data here analysed data consist of demographic information on the users and of transaction records of all registered accounts across 70 financial institutions in the UK. These transactions carry information on vendor, location and purpose of the transaction. Because of their structure and information content the MD data can be considered to be of OB-type (details on OB can be found in the Introduction). The MD dataset is unique in detail and provides a rare in-depth look into the daily financial behaviour of UK customers.

One drawback of the MD sample is that it consists of a self-selected group of MD application users. This has some implications on the variable distributions. Because we focus on specific users, as long as we do not draw conclusions for the whole UK population from our results, our analysis and findings are not affected.

The data consist of anonymised transaction records of 10,689 MD application users for 2017 and 2018. Using these information we categorise the transactions and aggregate income $X_{i,t}$ and expenditure $Y_{i,t}$, on a monthly basis, of any user i in month t. We initially focus on 2017. In Section 4.3 we include the 2018 data in our analysis.

The violin plots in Figure 4 depict the monthly distributions of income and expenditure in the analysed 2017 data, and on average income and expenses are balanced out. However, the heavy tails of both the income and expenditure distributions in time are noticeable. This is highlighted by the difference in the average income of £10, 293.29 and the median income of £4, 767.615. The same goes for expenditure, where the mean is £9, 923.63 and the median is £4, 750.55. As users utilise MD services to manage multiple accounts, the data consist of a self-selected group of customers with sufficient funds and capability to manage them. In comparison with the total UK population, this leads to upward shifted distributions in the MD sample. For example, the HM Revenue and Customs (2018)

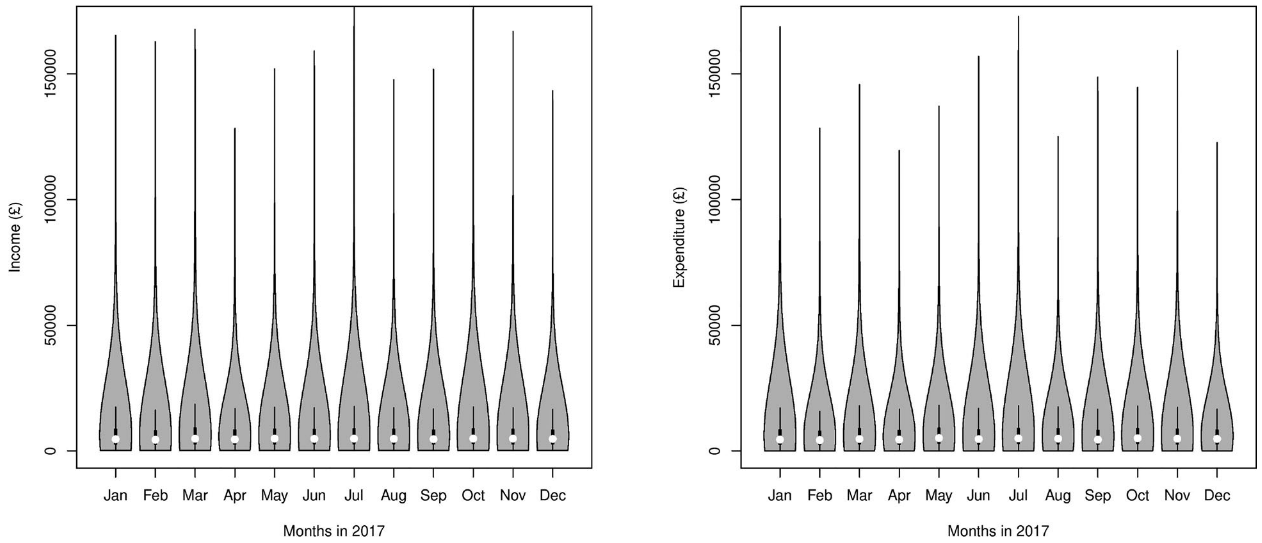

Figure 4. Violin plot of income (left) and expenditure (right) over 2017.

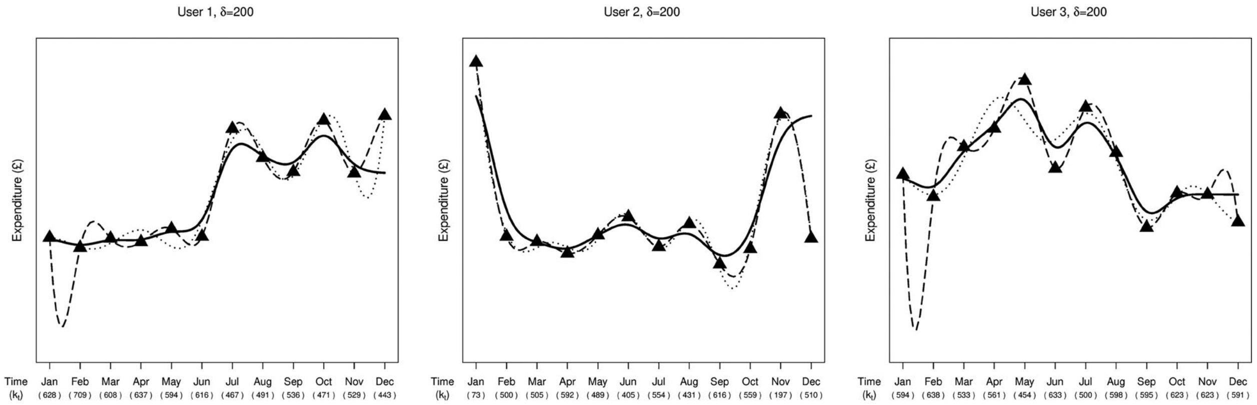

Figure 5. Smooth time-varying comparative mean expenditure: The time-varying comparative mean expenditure (7) is represented by the triangles (▲), and the smooth comparative mean expenditure is represented by solid (kernel, (9)), dashed (Gaussian process), and dotted (P-spline) lines. The *x*-axis denotes time, and the number of subjects in the window of comparison, k_t.

obtained a mean income of £2, 891.66 and median income of £2, 033.33 for 2017.

4.2. Time-varying comparative mean expenses

After having gained an overview of the MD data we next apply the time-varying comparative sample mean (7) with expenditure being the response ($Y_{i,t}$) and income being the covariate ($X_{i,t}$). We select three distinct users (reference subjects) with different financial behaviour from the data set, whose expenditure ($y_{0,t}$) can be seen in Figure 6, and whose income ($x_{0,t}$) we use as reference covariates the peer selection in (7). The peer selection is based on income and the expected value of expenditure, conditioned on the selected peer group is calculated as in (7). Using the the asymptotically optimal bandwidth for kernel density methods, given by DasGupta (2008, p. 531), we provide the smooth time-varying comparative sample mean (9). Then the results are compared to the response of the reference subject. In other words, we will be tracking how a user's expenses develop in comparison to peers, who have similar income.

Figure 5 shows the time-varying comparative mean expenses (7), and the smooth time-varying comparative mean expenditure (9) for three different users at $\delta = £200$; for privacy reasons, we are not allowed to display the y values nor on Figure 5 nor on any subject-specific fit presented below. For comparison, we also present in Figure 5 the smooth time-varying comparative mean expenditure that is learned by other two popular methods, namely Gaussian processes (Rasmussen & Williams, 2006) and P-splines (Eilers & Marx, 1996); the results from all smoothers are essentially equivalent. The value of δ was chosen to allow the peers to have very similar, but not identical income as the reference subject. In each panel of Figure 5 users with income in the range of ±£200 from the reference

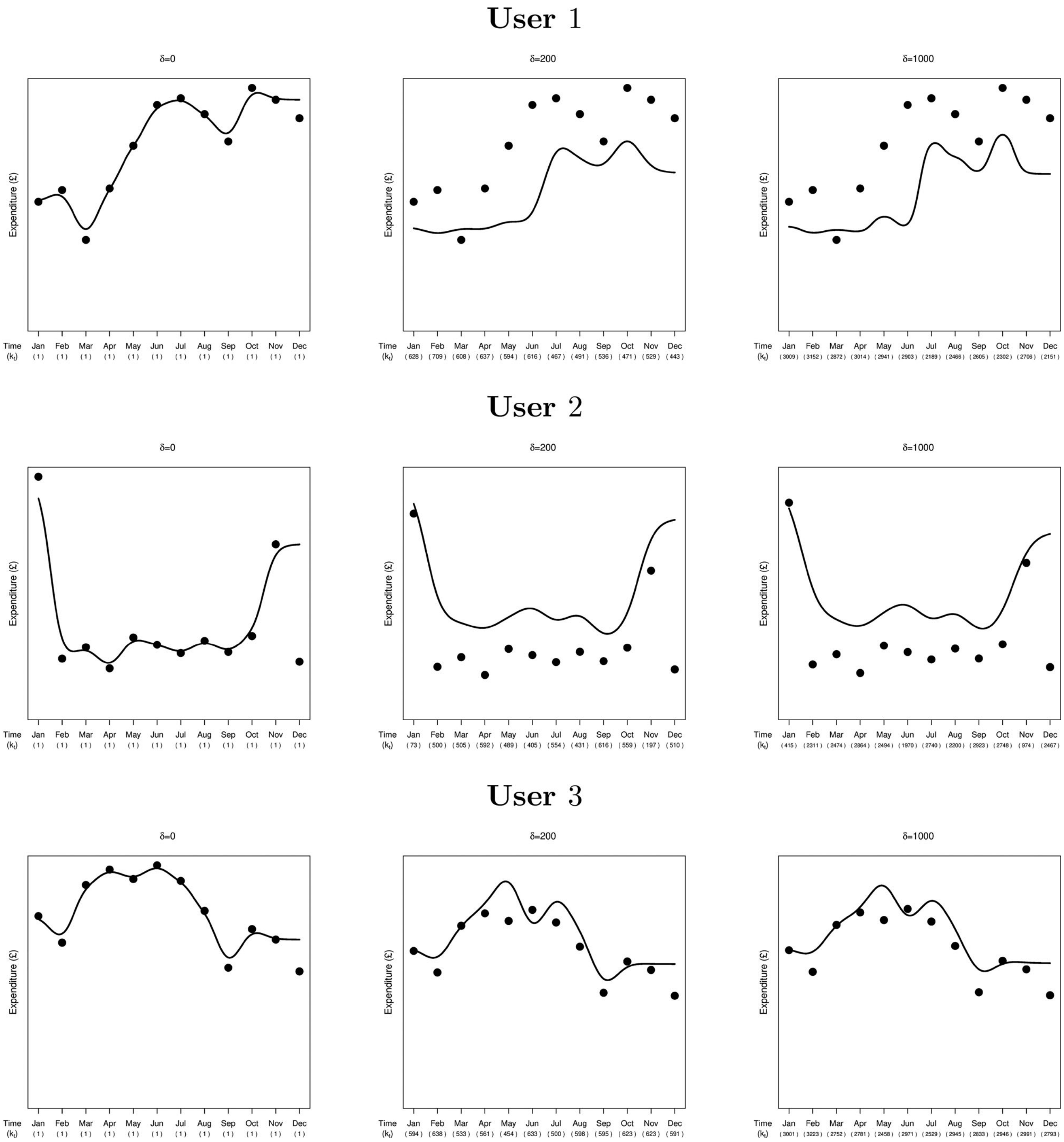

Figure 6. Time-varying comparative mean expenditure analysis: The smooth time-varying comparative mean expenditure is represented by the solid line; the dots (●) are the user expenses $y_{0,t}$. The x-axis denotes time, and the number of subjects in the window of comparison, k_t. Each column corresponds to a different window of comparison.

user's income are selected and the time-varying comparative mean expenditure (7) is given. As can be seen in Figure 5, smoothing provides a natural way to track the dynamics of the time-varying comparative mean expenditure over time.

In Figure 6, we contrast the smooth time-varying comparative mean expenditure (9) with the expenditure of the reference subjects for three different values of δ. To give an example of how to interpret the obtained results, we will focus on the case $\delta = £200$. The smooth time-varying comparative mean expenditure (9) represents the mean expenditure of users with income ±£200 in range of the reference user's income. The points in Figure 6 are the expenses of the reference users in each month of 2017. We can notice how the three reference users considerably differ in their spending behaviour. With the exception March, user 1 spends consistently more than the average of users with the same income ±£200. The opposite case is user 2, who tends to spend less than the average of users with similar income. Exceptions are January and November, when the user's spending is aligned with peers. For user 3 spending is lined up with that of the peer average, with the exceptional deviations in February, September and December of 2017.

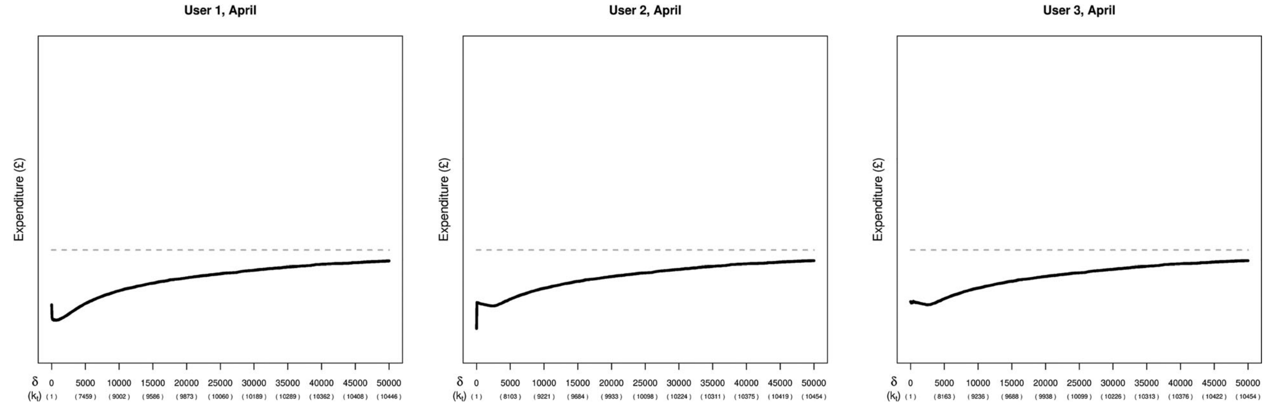

Figure 7. Comparative mean expenditure over all windows of comparison: The comparative mean expenditure (7) is represented by the black curve. The grey dotted curve represents the sample mean of expenditure in the given month. The *x*-axis shows the δ values and the number of subjects in the window of comparison, k_t.

Given the information in Figure 6, one can provide guidance on personal finance to the respective users. For illustration, because user 1 tends to spend consistently more money than peers with similar income, the user might be considered at risk of immediate or future financial problems. One can provide advice on how to reduce spending or occasionally send notifications to such users to remind them of their tendency to overspend. For user 2, who spends consistently less than peers with similar income, one can focus on advice on savings and investments. And last but not least, advisers can send user three warnings or recommendations when the user's spending behaviour starts to deviate from the peer average. The guidance also needs to take account the financial situation within the peer group, as even customers with similar income might be on average overspending or underspending. However, using the comparative mean value one is able to identify group behaviour of the comparison group, behaviour of reference subjects and detect customers with different behavioural patterns in comparison to peers.

Functions measuring the difference between the reference subject value and the time-varying comparative mean value are a possibility for identifying each of the three cases mentioned above. However, the usage of absolute measures requires separating cases when the reference subject's spending is above or below the time-varying comparative mean value so to distinguish between cases of overspending and underspending.

For a more complete picture of how the comparative mean expenditure changes with different values of δ, in Figure 7 we select a specific month and plot the resulting comparative mean expenditure (7) with the corresponding values of δ.

The interpretation of Figure 7 is as follows. The starting position of the comparative mean expenditure at $\delta = 0$ in respect to the total sample mean provides information on the reference subjects' position in the total distribution of expenditure. It is noticeable, that despite of the very different expense behaviour of our reference subjects, all three reference subjects spend less than the total sample on average. As the distribution of total expenses is heavy-tailed, most users have expenses below the total sample average.

The slope of the comparative mean expenditure as a function of δ tells us how much the reference subjects differ from peers with similar income. In absolute, the greater the slope at low values of δ, the more the specific user's response deviates from that of the average in the peer group. For large positive slopes the user is spending considerably less than peers with very similar income. The smaller a negative slope, the more the user spends in comparison to their closest peer groups. As we keep increasing δ more people are included in the window of comparison and the comparative mean expenditure approaches the total sample mean. To illustrate, we turn back to Figure 7. As discussed previously, we already know that user 1 tends to spend consistently more than peers with similar income. For April 2017, we can clearly see this in Figure 7 on the very steep negative slope around $\delta \approx 0$. Furthermore, for any $\delta < £5,000$ the user's expenditure, seen at $\delta = 0$, stays above mean peer expenditure. This means that user 1 spends more than the average of any peer group selected based on similar income of $\pm £5,000$ of the user's income. As mentioned previously, this can become or already be a financial problem for the user. The opposite case is user 2, who tends to spend less than peers. This is in line with the slope of the comparative mean expenditure at low δ values in Figure 7. A very small increase of δ around $\delta \approx 0$ increases the comparative mean expenditure considerably. This indicates that users with very similar income spend much more than user 2. As we keep increasing δ the comparative mean expenditure dips a bit, but stays above the reference subject's value. This means the user 2 spends less than any peer

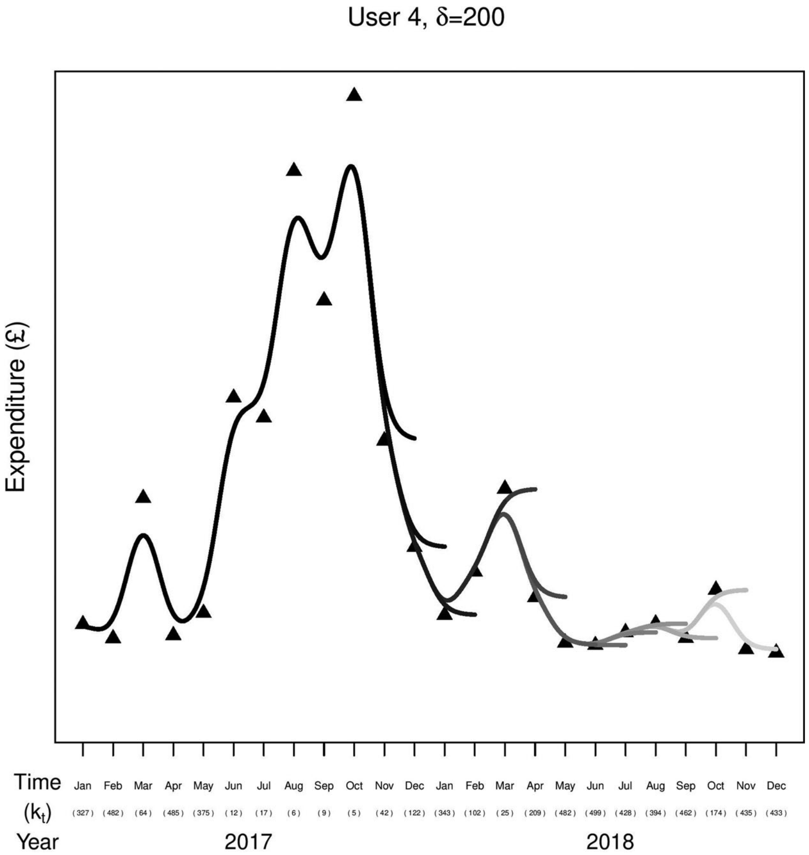

Figure 8. Real-time smooth time-varying comparative mean value analysis: Starting with January 2018 new data points are added and the smooth comparative mean expenditure (7) is fit. The lighter the curves the more data points are considered. The *x*-axis shows the number of subjects in the window of comparison k_t, and months between January 2017 and December 2018.

group selected using the window of comparison. Last but not least user 3 who spends roughly the same as the average of peers. We can see the comparative mean expenditure does not notably change at any delta of approximately $\delta<£2,000$. At around $\delta\approx£3,000$ the comparative mean value drops below the user's value. This indicates more customers with lower expenditure in the window of comparison. For values of approximately $\delta>£5,000$ this difference disappears and the comparative mean expenditure stays above the user's response.

A commonly used non-statistical approach for personal financial analysis are financial ratios (e.g., see Bae et al., 2005; Greninger et al., 1996; Melzer, 2017; Organisation for Economic Co-operation and Development (OECD), 2020) a review of which can be found in Harness et al. (2008). We present the expense to income ratios for users 1–3 in the Supplementary Material. However, we would like to point out that financial ratios were constructed for other purposes than for subject-to-group comparison pioneered in this paper.

4.3. Further empirical analysis

Here we provide further illustrations of the proposed approach, with a particular focus on monitoring:

1. How the estimates of smooth comparative mean expenses get updated once new data arrives?
2. Is there any major impact on the subject-specific fits as the database structure evolves over time?

In terms of (1), in fields such as Econometrics the ability of a method to be coherent over real-time—in the sense of not revising estimates, once new data arrives—is key, and it has been a subject of wide interest (see, for instance Orphanides and Van Norden (2002), and references therein). In terms of (2), the data analysed below are similar to that from Section 4.1, but with minor changes in transaction categorisation, and consisting of two years of data (2017–2018). As it can be seen from Figure 8 revisions are moderate except at the end of the observation period, which is mostly explained by the well-known boundary–bias challenges faced by the Nadaraya–Watson estimator (Wand, 1995, Section 5.5).

5. Discussion and closing remarks

Unlike methods for comparing population groups (e.g., *t*-test), subject-to-group comparison has received little attention in Operational Research and in Statistics, but as this paper puts forward the latter comparison is of the utmost importance in an OB setup. In this paper, we introduced the time-varying comparative mean value, which is a statistical method for learning about the dynamics of differences between the response of a reference subject and that of a comparison group defined by covariate truncation around the reference subject's covariates. The covariate truncation is controlled by a similarity variable (δ) set by the user. By varying δ one can extract information on the distribution of the response and the position of the reference subject in it. No parametric structure on the response and/or covariates is assumed. The method can be viewed as a time-varying truncated covariate regression model, and local polynomial regression is used to define its smooth version. A simulation study showed the estimators recover the true time-varying comparative mean value well.

As discussed in Section 2.4, the time-varying comparative mean has links to statistical concepts such as the F-barycenter, MES and SES, regression towards the mean and k-nearest neighbours. There are also links to regression methods for truncated random variables. Yet, unlike the latter regression methods which are designed for truncated responses, the time-varying comparative mean value deals with truncated covariates. To our knowledge the time-varying comparative mean value provides a first methodological approach to truncated regression on an interval centred around a reference covariate.

We showcase the usefulness and insights obtained by the time-varying comparative mean value on data provided by a UK service provider, MD, who motivated the development of the here proposed methods. Because for each customer in the sample of 10,689 customers, the data covers all transactions in 2017 and 2018 for all accounts registered by the customers across 70 financial institutions in the UK, the data can be considered to be of OB-type. In the analysis we compare the spending behaviour of selected reference customers with the spending behaviour of customers with similar income as the individual reference customers. Based on the diagnostics obtained by such analyses the financial services provider can offer advice and services specifically tailored for the financial needs of the reference customer. Because the insights obtained by the time-varying comparative mean value are also simple to read and communicate, when implemented into a financial services platform or a financial application, customers can contrast their own finances, e.g., expenses or savings, against customers they consider to be financial peers, e.g., people with on similar income or age.

Although we apply the time-varying comparative mean value to personal finance, due to its generality it can be used in other contexts, such as comparing one patient to patients with similar characteristics to identify patient-to-group differences.

To allow for a more granular definition of peers the time-varying comparative mean value can be defined using a window of comparison with multiple explanatory variables. While here we have focussed on a window of comparisons based on the single covariate framework, the extension of the proposed methods for multiple covariates is feasible and will be explored elsewhere.

Tracking how the dynamics of the comparative mean value may change when there are potential jumps or breaks in the data is a natural challenge that one may face in some applications. Versions of the proposed time-varying comparative value that resort to models for curves with jumps (e.g., Gijbels et al., 2004; Kang, 2020) or to wavelets (Abramovich et al., 2000)—rather than local polyonimial methods—may be more natural for those applications.

To identify customers at risk, it would also seem natural contrasting the reference subject against a high quantile of the response, such as expenses, say on a window of comparison based on subjects with similar covariates, such as income. This would lead to comparative quantile-based approaches, which would have links with quantile regression (Koenker, 2005). Specifically, the time-varying comparative quantile could be defined as

$$q_\delta = Q_{Y_t}\{p|X_t \in W_{\delta,t}(x_{0,t})\}, \quad 0<p<1.$$

This is another natural avenue for future research.

Acknowledgements

We thank the participants of the Financial Health Check Data Fest 2018, CMStatistics 2018–2019, Credit Scoring and Credit Control Conference XVI, and to partners from the Scottish Financial Risk Academy for discussions and comments, as well as to Data Lab (project no. 17472) and Inbest for funding, collaboration, and constant support. Special thanks goes to Money Dashboard for providing the data for the analysis. M. de Carvalho acknowledges funding from the FCT (Fundação para a Ciência e a Tecnologia, Portugal) through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019. Raffaella Calabrese acknowledges support from grant number ES/L011859/1, from The Business and Local Government Data Research Centre, funded by the Economic and Social Research Council.

Disclosure statement

There are no known conflicts of interests.

Funding

This research was supported by Fundação para a Ciencia e a Tecnologia, Portugal; Inbest.ai; The Business and Local Government Data Research Centre (Economic and Social Research Council); and The Data Lab.

ORCID

M. de Carvalho http://orcid.org/0000-0003-3248-6984
R. Calabrese http://orcid.org/0000-0002-0078-3151

References

Abramovich, F., Bailey, T. C., & Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *49*(1), 1–29. https://doi.org/10.1111/1467-9884.00216

Acharya, V., Pedersen, L. H., Philippon, T., & Richardson, M. P. (2010). *Measuring systemic risk* FRB of Cleveland Working Paper No. 10–02.

Acharya, V., Pedersen, L. H., Philippon, T., & Richardson, M. P. (2017). Measuring systemic risk. *Review of Financial Studies*, *30*(1), 2–47. https://doi.org/10.1093/rfs/hhw088

Bae, M. K., Hanna, S., & Baek, E. (2005). Before and after the economic crisis: Changes in financial ratios of the self-employed households. *Consumer Interests Annual*, *51*, 290–295.

Breidbach, C. F., Keating, B. W., & Lim, C. (2019). Fintech: Research directions to explore the digital transformation of financial service systems. *Journal of Service Theory and Practice*, *30*(1), 79–102. https://doi.org/10.1108/JSTP-08-2018-0185

Brodsky, L., & Oakes, L. (2017). Data sharing and open banking. McKinsey & Company. https://www.mckinsey.it/sites/default/files/data-sharing-and-open-banking.pdf

Buckley, R. P., Arner, D. W., Zetzsche, D. A., & Weber, R. H. (2020). The road to RegTech: The (astonishing) example of the European Union. *Journal of Banking Regulation*, *21*(1), 26–36. https://doi.org/10.1057/s41261-019-00104-1

Cai, J. J., Einmahl, J. H., Haan, L., & Zhou, C. (2015). Estimation of the marginal expected shortfall: The mean when a related variable is extreme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(2), 417–442. https://doi.org/10.1111/rssb.12069

Cheng, L., & Zagat, E. (2019). U.S. patent no. 10,453,152. U.S. Patent and Trademark Office.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer.

De Cnudde, S., Moeyersoms, J., Stankova, M., Tobback, E., Javaly, V., & Martens, D. (2019). What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance. *Journal of the Operational Research Society*, *70*(3), 353–363. https://doi.org/10.1080/01605682.2018.1434402

Efron, B., & Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, *94*(447), 824–834. https://doi.org/10.1080/01621459.1999.10474187

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–102. https://doi.org/10.1214/ss/1038425655

European Commission. (2018). *Implementation of PSD2 directive*. https://ec.europa.eu/info/law/payment-services-psd-2-directive-eu-2015-2366/implementation

Fan, J. (1996). *Local polynomial modelling and its applications*. Chapman & Hall.

Financial Conduct Authority. (2017). *Financial advice market review baseline report*. https://www.fca.org.uk/publication/research/famr-baseline-report.pdf

Gijbels, I., Hall, P., & Kneip, A. (2004). Interval and band estimation for curves with jumps. *Journal of Applied Probability*, *41*(A), 65–79. https://doi.org/10.1239/jap/1082552191

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of Management Information Systems*, *35*(1), 220–265. https://doi.org/10.1080/07421222.2018.1440766

Greninger, S. A., Hampton, V. L., Kitt, K. A., & Achacoso, J. A. (1996). Ratios and benchmarks for measuring the financial well-being of families and individuals. *Financial Services Review*, *5*(1), 57–70. https://doi.org/10.1016/S1057-0810(96)90027-X

Harness, N., Chatterjee, S., & Finke, M. (2008). Household financial ratios: A review of literature. *Journal of Personal Finance*, *6*(4), 77–97.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer.

He, Z., Huang, J., & Zhou, J. (2020). Open banking: Credit market competition when borrowers own the data [No. w28118]. National Bureau of Economic Research.

Hill, T., & Monticino, M. (1998). Constructions of random distributions via sequential barycenters. *The Annals of Statistics*, *26*(4), 1242–1253. https://doi.org/10.1214/aos/1024691241

HM Revenue and Customs. (2018). *Distribution of median and mean income and tax by age range and gender: 2017 to 2018*. Retrieved March 30, 2021 https://www.gov.uk/government/statistics/distribution-of-median-and-mean-income-and-tax-by-age-range-and-gender-2010-to-2011

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*. Wiley.

Kang, Y. (2020). Measuring timeliness of annual reports filing by jump additive models. *The Annals of Applied Statistics*, *14*(4), 1604–1621. https://doi.org/10.1214/20-AOAS1365

Koenker, R. (2005). *Quantile regression*. University Press.

Lee, I., & Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business Horizons*, *61*(1), 35–46. https://doi.org/10.1016/j.bushor.2017.09.003

Li, Q., & Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*, 485–512.

Melzer, B. T. (2017). Mortgage debt overhang: Reduced investment by homeowners at risk of default. *The Journal of Finance*, *72*(2), 575–612. https://doi.org/10.1111/jofi.12482

Mention, A. L. (2019). The future of fintech. *Research-Technology Management*, *62*(4), 59–63. https://doi.org/10.1080/08956308.2019.1613123

Nicoletti, B., Nicoletti, W., & Weis. (2017). *Future of FinTech*. Palgrave Macmillan.

Omarini, A. E. (2018). Banks and FinTechs: How to develop a digital open banking approach for the bank's future. *International Business Research*, *11* (9), 23–36. https://doi.org/10.5539/ibr.v11n9p23

Organisation for Economic Co-operation and Development. (2020). *Housing costs over income*. Retrieved 30 March, 2021 https://www.oecd.org/els/family/HC1-2-Housing-costs-over-income.pdf

Orphanides, A., & Van Norden, S. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, *84*(4), 569–583. https://doi.org/10.1162/003465302760556422

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, *74*, 26–39. https://doi.org/10.1016/j.asoc.2018.10.004

Racine, J. (2001). Bias-corrected kernel regression. *Journal of Quantitative Economics*, *17*, 25–42.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Ruiz, J., & Navarro, J. (1996). Characterizations based on conditional expectations of the doubled truncated distribution. *Annals of the Institute of Statistical Mathematics*, *48*(3), 563–572. https://doi.org/10.1007/BF00050855

Shen, P. S. (2010). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics*, *62*(5), 835–853. https://doi.org/10.1007/s10463-008-0192-2

Stranieri, A., McInnes, A. N., Hashmi, M., Sahama, T. (2021). Open banking and electronic health records. *Proceedings of the Australasian Computer Science Week Multiconference*, 1–4.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *38*(3), 290–295. https://doi.org/10.1111/j.2517-6161.1976.tb01597.x

Wand, M. P. (1995). *Kernel smoothing*. Chapman & Hall.

Wang, H., Ma, S., Dai, H. N., Imran, M., & Wang, T. (2020). Blockchain-based data privacy management with nudge theory in open banking. *Future Generation Computer Systems*, *110*, 812–823. https://doi.org/10.1016/j.future.2019.09.010

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, *26*(4), 359–372.

Ying, Z., Yu, W., Zhao, Z., & Zheng, M. (2019). Regression analysis of doubly truncated data. *Journal of the American Statistical Association*, *115*(530), 1–25.

Zhu, W., Liu, B., Lu, Z., & Yu, Y. (2020). A DEALG methodology for prediction of effective customers of internet financial loan products. *Journal of the Operational Research Society*, 1–9, 1033–1041.