Modelling corporate bank accounts

May 2021

Abstract

We discuss the modelling of corporate bank accounts using a proprietary dataset. We thus offer a principled treatment of a genuine industrial problem. The corporate bank accounts in our study constitute spare, irregularly-spaced time series that may take both positive and negative values. We thus builds on previous models where the underlying is real-valued. We describe an intra-monthly effect identified by practitioners whereby account uncertainty is typically lowest at the beginning and end of each month and highest in the middle. However, our theory also allows for the opposite effect to occur. In-sample applications demonstrate the statistical significance of the hypothesised monthly effect. Out-of-sample forecasting applications offer a 9% improvement compared to a standard SARIMA approach.

Keywords: Corporate Bank Accounts; Fin Tech; Forecasting Applications; Machine Learning
JEL classification: C22; E3; G1; G3

1 Introduction

In this paper we discuss the problem of forecasting daily cash balances of corporate bank accounts. This is an area that has been under-explored academically (Griguta et al., 2021). This paper thus provides a principled treatment of a genuine industrial problem. The original motivation was to improve forecast accuracy to enable the optimal resource allocation to be achieved across higher-yield accounts. The model presented here thus enables the development of important foundational levels of understanding on the way to achieving this goal. This is significant for two reasons. Firstly, there is conceptual value in providing a model where account uncertainty is explicitly tied to monthly business-cycle fluctuations. Secondly, there is additional value in coherently quantifying account uncertainty out-of-sample. Here, in the spirit of machine learning, we concentrate upon a simple forecasting comparison. However, our approach could ultimately be used to develop new forms of account insurance.

The corporate bank account data is extracted from bank statements collated through the SWIFT network. This is in itself quite an involved process. The resulting series have several missing values and might be reported only sparsely for some bank accounts. Moreover, some

of the corporate bank accounts considered only had data available for 6-9 months. Whilst far from ideal this nonetheless emphasises the practical nature of the problem at hand. Further, ignoring these data limitations is a luxury we could ill afford. An exploratory analysis of this data is contained in Griguta et al. (2021). Summary statistics and other qualitative features of this data are discussed in Section 3.

In line with pragmatic aspects of financial model construction (Cont and Tankov, 2004) a model in which the volatility is subject to monthly fluctuations around a constant drift gives a parsimonious way of modelling realistic dynamical behaviour. Further, we adopt a Gaussian model, rather than the standard log-Gaussian formulation, to account for the fact that corporate bank accounts are not constrained to be positive. This leads to an elegant generalisation of Bachelier's classical model (Bachelier, 1900; Bouchaud and Potters, 2003) and follows recent models that have been constructed in order to allow the underlying to take negative values (Carr and Torricelli, 2020).

In this paper we provide new stochastic modelling and new ways of conceptualising corporate bank accounts. The importance of our contribution is twofold. Firstly, we use an SDE model to quantify a proposed monthly effect identified by practitioners. This has practical significance in that typically the uncertainties associated with managed accounts are highest towards the middle of the month and lowest at the month ends. However, our theory also allows for the reverse effect whereby the uncertainty is lowest at the middle of the month and highest at the month ends. The background to this study is ultimately very rich. Section 3 demonstrates the empirical significance of this hypothesised monthly effect. Secondly, we use this SDE model to derive principled out-of-sample forecasts using this model. Alongside analytical tractability this approach is shown to provide some improvement in forecasting results with respect to a standard SARIMA approach in Section 4.

The layout of this paper is as follows. Section 2 discusses the stochastic modelling of corporate bank account balances. Section 3 conducts an empirical test for monthly effects within the data. Section 4 outlines an out-of-sample forecasting application. Section 5 concludes and discusses the opportunities for further research.

2 Stochastic modelling of corporate bank account balances

In contrast to most financial and economic time series the corporate bank accounts considered in this paper are not constrained to be positive. Further, there has been increased recent interest in models where the underlying can take negative values (Carr and Torricelli, 2020). As such we consider the following modification to Bachelier's classical Gaussian model (Bachelier, 1900; Bouchaud and Potters, 2003). Let

$$dP(t) = \mu dt + \sigma(t) dW_t, \tag{1}$$

where P(t) in (1) denotes the balance. In industrial applications, and in consultation with practitioners, it is natural to consider that the uncertainty around $P(t)$ is lowest at either the beginning ($t = t_1$) or the end ($t = t_2$) of the month. This monthly effect may reflect conscious management of the accounts in question. However, the following parameterisation allows us to model the reverse effect where, in contrast, the uncertainty is highest at the month end and lowest in the middle. Thus, we choose the following trigonometric form for $\sigma^2(t)$:

$$\sigma^2(t) = \sigma_0^2 + (\sigma_{Mid}^2 - \sigma_0^2)\sin^2\left(\frac{\pi(t - t_L)}{t_U - t_L}\right), \tag{2}$$

where σ_0^2 describes the level of uncertainty that occurs at the end of each month and σ_{Mid}^2 is the level of uncertainty associated with the middle of each month. The values t_L and t_U correspond to the dates of the month open and the month close respectively. Set up in this way $\sigma^{-2}(t)$ also satisfies an integrability constraint laid out in Bingham and Kiesel (2004), Chapter 6.2. Our original motivation was thus to extend recent empirical forecasting applications based around the theory of binary options (see e.g. Taleb, 2018; Fry and Burke, 2020). However, liquidity issues mean that it may be difficult to apply standard options-pricing arguments in the case of corporate bank accounts. See e.g. the related discussion in Battauz et al. (2012).

In the absence of a monthly effect $\sigma_{Mid}^2 = \sigma_0^2$ and the distribution of the difference between successive balances can be written as

$$P_{t_2} - P_{t_1}|P_{t_1} \sim N(\mu(t_2 - t_1), \sigma_0^2(t_2 - t_1)). \tag{3}$$

Equation (3) thus shows that in the simplified case of $\sigma_{Mid}^2 = \sigma_0^2$ the distribution of successive Treasury balances just depends on the time difference. However, in the full monthly effects model $\sigma_{Mid}^2 \neq \sigma_0^2$ this distribution can be written as

$$P_{t_2} - P_{t_1}|P_{t_1} \sim N\left(\mu(t_2 - t_1), \left(\frac{\sigma_0^2 + \sigma_{Mid}^2}{2}\right)(t_2 - t_1) + M_t\right), \tag{4}$$

where M_t is a monthly adjustment given by

$$\begin{aligned} M_t &= \left(\frac{t_{U_2} - t_{L_2}}{4\pi}\right)(\sigma_0^2 - \sigma_{Mid}^2)\left[\sin\left(\frac{2\pi(t_2 - t_{L_2})}{t_{U_2} - t_{L_2}}\right)\right] \\ &- \left(\frac{t_{U_1} - t_{L_1}}{4\pi}\right)(\sigma_0^2 - \sigma_{Mid}^2)\left[\sin\left(\frac{2\pi(t_1 - t_{L_1})}{t_{U_1} - t_{L_1}}\right)\right]. \end{aligned} \tag{5}$$

Note that in equation (5) the values L_1, U_1 and L_2 and U_2 are the monthly open and monthly close dates corresponding to times t_1 and times t_2 respectively.

Motivated by out-of-sample forecasting applications in Section 4 we have that

$$P_t|P_0 \sim N\left(P_0 + \mu t, \int_0^t \sigma^2(u)du\right). \tag{6}$$

From equation (6) an out-of-sample forecast can be calculated as

$$E[P_t|P_0] = P_0 + \mu t. \tag{7}$$

An associated $100(1-\alpha)\%$ confidence interval can be constructed using

$$\text{Condidence Interval} = P_0 + \mu t \pm \Phi^{-1}(1-\alpha/2)\sqrt{\int_0^t \sigma^2(u)du}, \tag{8}$$

where $\Phi^{-1}(\cdot)$ denotes the inverse CDF of a $N(0,1)$ random variable.

3 In-sample application: model calibration via maximum likelihood

The data for this study constitutes a proprietary dataset of 19 corporate bank accounts. Bank statements issued for corporate users are similar in format to retail bank statements. They contain a header detailing the account name, identification code, date and time of issue and the opening and closing balance. These balances can take both positive and negative values and constitute an irregularly-spaced time series since account data can be reported according to different granularities and conventions. Some accounts are reported every calendar day. Other accounts are only reported on certain calendar days. Following standard practice (see e.g. Fry and Serbera, 2020) summary statistics for this data are shown below in Table 1. A graphical illustration of the seasonal patterns apparent within this data is shown below in Figure 1.

From equations (2-3) Table 2 tests the null hypothesis of no monthly effect ($\sigma^2_{Mid} = \sigma^2_0$) against a general alternative. Likelihood ratio tests in Table 2 thus present evidence of a monthly effect in 17 of the 19 series sampled.

4 Out-of-sample forecasting application: SDE approach versus machine learning

In this section we consider an out of sample forecasting application where we compare our binary options approach with a standard SARIMA forecasting approach (Brockwell and Davis, 2016) that has previously been considered in financial applications (see e.g. Saz, 2011).

Following the standard machine-learning approach (see e.g. Bishop, 2006) we split the data into a training phase shown in Section 4 and an out-of-sample test phase detailed in this section. Forecasted values from the SDE model in Section 2 are constructed using equation (7). As discussed in Section 3 the parameters μ, σ^2_0 and σ^2_{Mid} are estimated via maximum likelihood. We compare these predictions against the forecast results obtained by an SARIMA $(1,0,1)\times(1,1,1)_{22}$ model which, as discussed in Griguta et al. (2021), gives a crude way account-

Series	Mean	Median	Max	Min	St. Dev	Skewness	Kurtosis
1	-390.17	0.000	16050.05	-1.7123.19	30121.28	-0.907	27.005
2	30520.56	-1409.09	26574010	-21787900	3481265	1.574	26.189
3	1732.81	0.000	105307.3	-72523.11	20218.82	0.785	9.826
4	-23927.672	162.830	188617.760	-8397310.610	464917.717	-17.842	321.775
5	19603.78	83669.57	960857.3	-258995	391535.3	-2.789	14.800
6	2073.56	0.000	27023460	-27031130	4919832	0.014	14.839
7	11504.75	52553.33	1556430	-2229659	266323.6	-2.434	28.629
8	-18745.28	179280.7	2815114	-6940000	1322025	-2.849	12.383
9	6257.07	61015.19	488575.1	-1774749	254472.3	-4.057	23.528
10	2459.64	209744.7	2188750	-4283015	941955	-1.788	7.550
11	-30905	-1329.57	26580800	-21986100	3490106	1.564	26.057
12	22400.86	1035194	24816200	-24719620	5263607	-1.488	10.303
13	-11760.33	0.000	602178.2	-4429266	305997.3	-9.088	121.508
14	9093.870	5992.033	2867611.827	-322012.228	161490.820	16.439	293.193
15	3235.77	25622.24	370208.7	-1649458	197590.7	-4.262	27.796
16	-13480.54	-158686.1	126786200	-117894800	9661625	1.472	120.479
17	10204.69	30450.35	2879489	-1633552	257601.9	2.342	57.533
18	-515233.7	-1747489	97346850	-57315510	17625760	1.792	8.898
19	15574.15	13911.69	2291525	-2727358	583854.5	-0.668	9.834

Table 1: Summary statistics for the first differences of the corporate bank accounts.

ing for monthly effects. Thus, having calibrated models against the training data $y_1, y_2, \ldots, y_n$ we construct forecast values $F_{n+1}, F_{n+2}, ..., F_{n+m}$, where m denotes the number of observations in the test phase, and compare with the historically observed values $y_{n+1}, y_{n+2}, \ldots, y_{n+m}$. Following Griguta et al. (2021) the forecasting performance of the two models is compared using the Normalized Root Mean Square Error (NRMSE) and the Symmetric in Mean Absolute Percentage Error (SMAPE) shown in equation (9):

$$\begin{aligned} \text{NRMSE} &= \sqrt{\frac{\sum_{i=n+1}^{n+m}(F_i - y_i)^2}{m\sum_{i=1}^{n}(y_i - \overline{y})^2}} \\ \text{SMAPE} &= \frac{100\%}{m}\sum_{i=n+1}^{n+m}\frac{2|F_i - y_i|}{(|F_i| + |y_i|)}. \end{aligned} \tag{9}$$

Out-of-sample forecasting results obtained for the period August 17th-September 15th 2020 are shown below in Tables 3-4. Results show that our proposed SDE approach performs marginally worse than the SARIMA model according to the NRMSE metric. However, using the SMAPE criterion, our proposed SDE approach offers a significant 9% improvement over the SARIMA model.

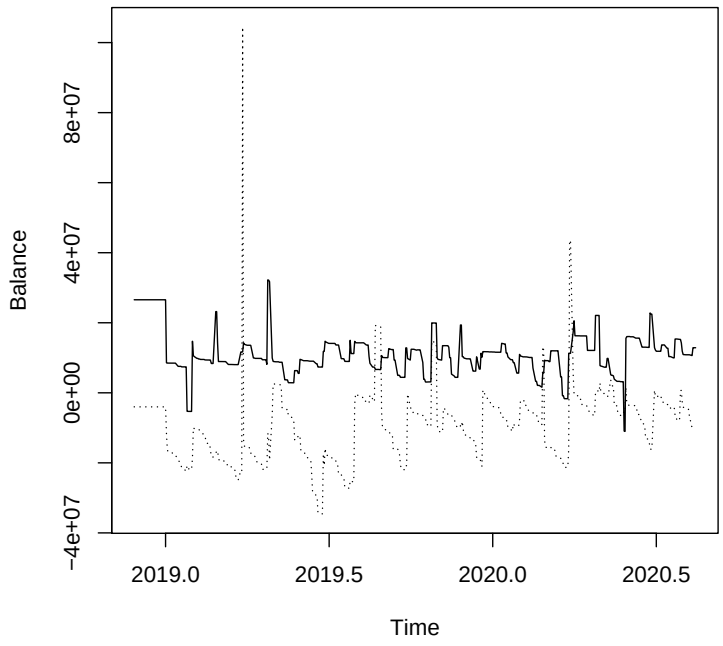

Figure 1: Exploratory plot showing approximate monthly seasonal effects in corporate bank account data November 26th 2018-August 14th 2020. Solid line: Series 2. Dashed line: Series 16.

5 Conclusions and further work

This paper provides a mathematical solution to a serious practical problem that has been under-explored academically – namely, stochastic modelling for quantifying uncertainty in corporate bank accounts. The practical nature of the problem at hand means it is difficult to forecast corporate bank accounts using conventional methods (Griguta et al., 2021). Particular problems include the spare and irregular nature of the time series involved – here solved using a continuous-time SDE model.

The importance of our contribution is two-fold. Firstly, we quantify a monthly effect identified by practitioners. usually account uncertainties are highest in the middle of the month and lowest at the month ends. However, our theory also allows for the reverse effect to occur. Secondly, we use an SDE model to conduct forecasts and, more importantly, to quantify the associated uncertainty. This is given further significance given the apparent failure of conventional methods (Griguta et al., 2021).Our work complements recent (Carr and Torricelli, 2020) and classical (Bachelier, 1900) mathematical finance models where the underlying can take negative values. Likelihood ratio tests present statistical evidence of an intra-monthly effect within corporate bank accounts. Formulae for point estimates and confidence intervals are derived

Series	Dates	No monthly-effects model	Monthly-effects model	χ^2-likelihood ratio ratio statistic	p-value
1	4/3/2020-14/8/2020	-1384.371	-1348.786	71.170	0.000
2	26/11/2018-14/8/2020	-7410.167	-7290.121	240.092	0.000
3	28/8/2019-14/8/2020	-2860.7	-2860.691	0.018	0.893
4	13/5/2019-14/8/2020	-4616.187	-4578.731	74.912	0.000
5	21/6/2019-14/8/2020	-4303.218	-4276.015	54.406	0.000
6	23/12/2019-14/8/2020	-2834.313	-2833.514	1.598	0.206
7	1/5/2019-14/8/2020	-4710.041	-4649.866	120.35	0.000
8	1/5/2019-14/8/2020	-5251.91	-5215.621	72.578	0.000
9	12/3/2019-14/8/2020	-5184.569	-5125.626	117.886	0.000
10	12/3/2019-14/8/2020	-5673.417	-5645.896	55.042	0.000
11	26/11/2018-14/8/2020	-7410.848	-7291.059	239.578	0.000
12	5/6/2019-14/8/2020	-5289.657	-5242.693	93.928	0.000
13	14/3/2019-14/8/2020	-5246.01	-5152.071	187.878	0.000
14	1/5/2019-14/8/2020	-4552.828	-4504.392	96.872	0.000
15	5/6/2019-14/8/2020	-4269.556	-4192.679	153.754	0.000
16	26/11/2018-14/8/2020	-7899.665	-7772.643	254.044	0.000
17	13/5/2019-14/8/2020	-4593.06	-4501.193	183.734	0.000
18	27/12/2018-14/8/2020	-7720.585	-7697.992	45.186	0.000
19	5/2/2020-14/8/2020	-2022.174	-2005.406	33.536	0.000

Table 2: Likelihood ratio tests of the null hypothesis of no monthly effect ($\sigma_0^2 = \sigma_{Mid}^2$) against a general alternative.

for out-of-sample forecasts. Our approach is shown to out-perform standard SARIMA models in out-of-sample forecasting tasks by up to around 9%. Results suggest new forms of account insurance and new derivatives contracts may ultimately be possible.

Acknowledgements

This work was funded by the Innovate UK, Knowledge Partnership 1025330 between Access Pay and Manchester Metropolitan University. The authors would like to thank Access Pay for their contribution and to acknowledge helpful and supportive comments from an anonymous reviewer. The usual disclaimer applies.

References

[1] Bachelier, L. 1900. Théorie de la spéculation. Ann. de l'École Norm. Supér., 1900, 3, 21-86.

[2] Battauz, A., De Donno, M., Sbuelz, A. 2012. Real options with a double continuation region. Quant. Finance, 12, 465-475.

Series	SDE model	SARIMA model
1	0.214	0.078
2	0.041	0.046
3	0.025	0.011
4	0.003	0.017
5	0.038	0.012
6	0.224	0.214
7	0.398	0.384
8	0.085	0.154
9	0.070	0.023
10	0.102	0.038
11	0.041	0.045
12	0.029	0.057
13	0.005	0.025
14	0.009	0.037
15	0.038	0.031
16	0.150	0.043
17	0.046	0.039
18	0.051	0.006
19	0.072	0.080
Mean	**0.086**	**0.063**

Table 3: Out-of-sample forecasting comparison using NRMSE: Monthly effects SDE model versus SARIMA.

[3] Bingham, N. H., Kiesel, R. 2004. Risk-neutral valuation, second ed.Springer.

[4] Bishop, C. M. 2006. Pattern recognition and machine learning. Springer, Singapore.

[5] Bouchaud, J-P., Potters, M. 2003. Theory of financial risk and derivative pricing: From statistical physics to risk management, second ed. Cambridge University Press, Cambridge.

[6] Brockwell, P. J., Davis, R. A. 2016. Introduction to time series and forecasting, third ed. Springer.

[7] Carr, P., Torricelli, L. 2020. Additive logistic processes in option prices. Preprint.

[8] Cont, R., Tankov, P. 2004. Financial modelling with jump processes. Chapman and Hall/CRC.

[9] Fry, J. M., Burke, M. 2020. An options-pricing approach to election prediction. Quant. Finance, 20, 1583-1589.

[10] Fry, J., Serbera, J-P. 2020. Quantifying the sustainability of Bitcoin and Blockchain. J. of Enterp. Inf. Manag., 33, 1379-1394.

Series	SDE model	SARIMA model
1	183.228	172.107
2	30.695	27.579
3	7.694	3.291
4	75.835	192.749
5	11.152	3.954
6	194.890	193.556
7	107.607	108.066
8	105.333	63.842
9	15.492	3.584
10	24.739	7.982
11	32.136	25.645
12	79.734	159.823
13	4.384	21.918
14	1.472	7.638
15	51.514	73.371
16	134.291	126.672
17	32.614	28.952
18	11.845	1.240
19	10.945	16.103
Mean	**58.716**	**67.372**

Table 4: Out-of-sample forecasting comparison using SMAPE: Monthly effects SDE model versus SARIMA.

[11] Griguta, V-M., Gerber, L., Slater-Petty, H., Crockett, K., Fry, J. M. 2021. Automated data processing of bank statements for cash balance forecasting. In Proceedings of Computing Conference 2021 (forthcoming)

[12] Saz, G. 2011. The efficacy of SARIMA models for forecasting inflation rates in developing countries: The case for Turkey. Int. Res. J. of Financ. and Bank., 62, 111-142.

[13] Taleb, N. N. 2018. Election predictions as martingales: An arbitrage approach. Quant. Finance, 2018, 18, 1-5.