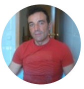
Empirical analysis on daily cash flow time series and its implications for forecasting

Article · November 2016

4 authors, including:

Francisco Salas Molina
Hilaturas Ferre, S.A.



Juan A. Rodríguez-Aguilar
Spanish National Research Council

Empirical analysis of daily cash flow time series and its implications for forecasting

Francisco Salas-Molina*[1], Juan A. Rodríguez-Aguilar[2], Joan Serrà[3], and Francisco J. Martin[4]

[1] *Hilaturas Ferre, S.A., Les Molines, 2, 03450 Banyeres de Mariola, Alicante, Spain*
[2] *IIIA-CSIC, Campus UAB, 08913 Cerdanyola, Catalonia, Spain*
[3] *Telefonica Research, Pl. Ernest Lluch, 5, 08019 Barcelona, Catalonia, Spain*
[4] *BigML, Inc, 2851 NW 9th Suite, Conifer Plaza Building, Corvallis, OR 97330, US*

November 17, 2016

Abstract

Cash management models determine policies based either on the statistical properties of daily cash flow or on forecasts. Usual assumptions on the statistical properties of daily cash flow include normality, independence and stationarity. Surprisingly, little empirical evidence confirming these assumptions has been provided. In this work, we provide a comprehensive study on 54 real-world daily cash flow data sets, which we also make publicly available. Apart from the previous assumptions, we also consider linearity, meaning that cash flow is proportional to a particular explanatory variable, and we propose a new cross-validated test for time series non-linearity. We further analyze the implications of all aforementioned assumptions for forecasting, showing that: (i) the usual assumption of normality, independence and stationarity hardly appear; (ii) non-linearity is often relevant for forecasting; and (iii) common data transformations such as outlier treatment and Box-Cox have little impact on linearity and normality. Our results highlight the utility of non-linear models as a justifiable alternative for time series forecasting.

1 Introduction

Cash management is concerned with the efficient use of a company's cash and short-term investments such as marketable securities. The focus is placed on maintaining the amount of available cash as low as possible, while still keeping a company operating efficiently. In addition, companies may place idle cash in short-term investments (Ross et al., 2002). Then, the cash management problem can be viewed as a trade-off between holding and transaction costs. If a company tries to keep balances too low, holding cost will be reduced, but undesirable situations of shortage will force to sell available marketable securities,

*Corresponding author. E-mail addresses: *francisco.salas@hifesa.com, martin@bigml.com, jar@iiia.csic.es, joan.serra@telefonica.com, arcos@iiia.csic.es*

hence increasing transaction costs. In contrast, if the balance is too high, low trading costs will be produced due to unexpected cash flow, but the company will carry high holding costs because no interest is earned on cash. Therefore, there is a target cash balance which each company must optimize according to the particular characteristics of its cash flows. An example of a cash flow time series is shown in Figure 1. This time series is apparently random, with values wandering around zero with some specific deviation, but perhaps following a hidden pattern.

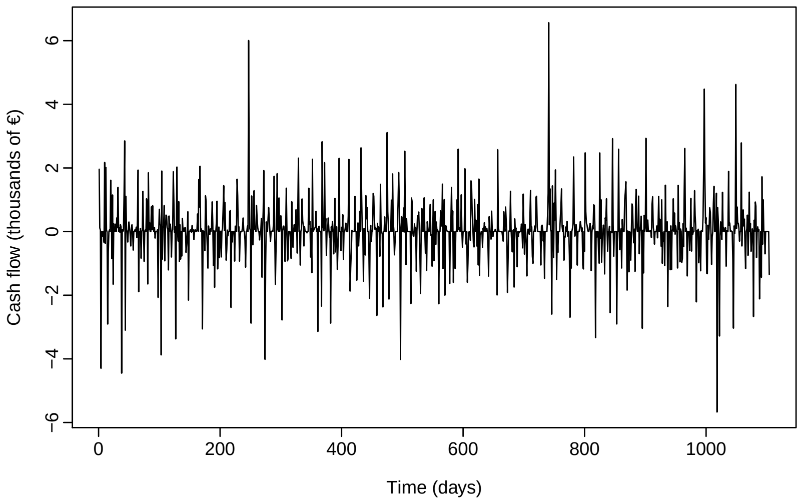


Figure 1: Example of a cash flow time series.

Since Baumol (1952), a number of cash management models have been proposed to control cash balances. These models are based either on the specific statistical properties of cash balances or on cash flow forecasts. A comprehensive review of models, from the first proposals to the most recent contributions, can be found in Gregory (1976); Srinivasan and Kim (1986), and da Costa Moraes et al. (2015). Most of them are based on assuming a given probability distribution for cash flows such as: (i) a random walk in the form of independent Bernouilli trials as in Miller and Orr (1966); (ii) a Wiener process as in Constantinides and Richard (1978); Premachandra (2004), and Baccarin (2009); (iii) a double exponential distribution as in Penttinen (1991). From these and other works, we observe that common assumptions on the statistical properties of cash flow time series include:

- Normality: cash flows follow a Gaussian distribution with observations symmetrically centered around the mean, and with finite variance.
- Independence: subsequent cash flows are not correlated in time with each other, meaning that the occurrence of past cash flows does not affect the probability of occurrence of the next ones.
- Stationarity: the probability distribution of cash flows does not change over time and, consequently, its statistical properties such as the mean and variance remain stable.

- Linearity: cash flows are proportional either to another (external) explanatory variable or to a combination of (external) explanatory variables.

Surprisingly, little and/or contradictory empirical evidence on these assumptions has been provided besides individual cases through time. Early on, negative normality tests were reported in Homonoff and Mullins (1975) for the times series samples of a manufacturing company. Contrastingly, later on, Emery (1981) reported normally distributed cash flow, after data transformation, for two out of three companies, and a small serial dependence for all of them. Pindado and Vico (1996) provided negative normality and independence results on 36 companies, but considering daily cash flow for only a single month. Previous works also reported day-of-week and day-of-month effects on cash flows, in line with the works of Stone and Wood (1977); Stone and Miller (1981); Miller and Stone (1985), and Stone and Miller (1987). Recently, Gormley and Meade (2007) described the time series from a multinational company with a non-normal distribution and serial dependence.

We consider that the evidence derived from these works is insufficient due to: (i) the disagreement between the conclusions of some of the works; (ii) the limited number of companies analyzed; and (iii) the short time range of the observations. Moreover, none of the previous works considered the presence of non-linear patterns for forecasting purposes. In this work, we provide an analysis of the statistical properties of 54 real cash flow data sets from small and medium companies in Spain, with annual revenue up to €10 million each. To the best of our knowledge, this is the most comprehensive empirical study on daily cash flow so far. We base this statement on the range of statistical properties considered, and on both the number and length of the data sets, which amount to 58005 observations in total, with a minimum, average and maximum time range of 170, 737, 1508 working days, respectively. A further contribution of the present work is to make all the aforementioned data publicly available online[1]. Finally, from a forecasting perspective, we also aim to identify the family of forecasters that best accommodate to our cash flow time series data sets. To this end, we propose a new and simple cross-validated test for non-linearity that provides further knowledge to cash managers in their search for better forecasting models.

Our results show the unlikely occurrence of normality, independence and stationarity. These results are consistent with the cited reports of Homonoff and Mullins (1975), based on only one time series, and Pindado and Vico (1996), based on a very short time range, raising doubts about the claim of independence. However, we also report that normality could not be achieved through removing outliers, contrary to what was reported by Emery (1981), based on only three time series. Thus, we consider that our results provide stronger evidence against normality, independence ans stationarity than previous works. In addition, our research departs from the cited works by considering two possible solutions to normality, independence and stationarity, namely, the Box-Cox transformation (Box and Cox, 1964), and time series differencing (Makridakis et al., 2008). However, we find that little benefit is derived from them when applied to our data sets. Our analysis also confirms the influence of seasonality as suggested in Miller and Stone (1985) and Stone and Miller (1987).

[1] `http://www.iiia.csic.es/~jar/54datasets3.csv`

When studying the properties of a time series, there is always an underlying question about data transformation. Is it always possible to achieve a Gaussian, noise-free and linear time series through data transformations? We here rely both on common statistical tests and on our novel non-linearity test to answer this question and we find that: (i) outlier treatment and Box-Cox transformation are not enough to achieve normality; (ii) outlier treatment produces mixed results in terms of noise reduction and information loss; (iii) outlier treatment and Box-Cox transformations do not produce linearity. These results suggest that non-linear models conform a justifiable alternative for cash flow time series forecasting, beyond the current conjectures of the literature.

The remaining of the paper is organized as follows. In Section 2, we provide a statistical summary of 54 real cash flow data sets including normality, independence and stationarity. In Section 3, we propose a new cross-validated test for non-linearity based on the comparison of a linear model and a non-linear model. Later, we present in Section 4 detailed results on the impact of data transformations on linearity. Finally, we provide some concluding remarks in Section 5.

2 Data summary

The data set contains daily cash flows from 54 different small and medium companies in Spain with annual revenue up to €10 million each. This data set covers a date range of about 8 years and is available online. An instance in the data set contains the following fields or columns:

- Date: standardized YYYY-MM-DD dates from 2009-01-01 to 2016-28-08.
- Company: company identifier from 1 to 54.
- NetCF: daily net cash flow in thousands of €.
- DayMonth: categorical variable with the day of the month from 1 to 31.
- DayWeek: categorical variable with the day of the week from 1 (Monday) to 7 (Sunday).

Table 1 shows the statistical summary of daily net cash flow on non-holidays, grouped by company. Due to the size of companies, almost 30% of the companies in our data set present more than 25% of null cash flow observations. This fact implies a first and important characteristic for forecasting: a null cash flow prediction will be right at least 25% of the times for this group of data sets. Therefore, two good baseline forecasting models for comparative purposes would be an *always-predict-null* or an *always-predict-mean* forecaster.

In addition, the average net cash flow shows that a high percentage of companies present either positive or negative drift with the exception of companies 5 and 28. High positive kurtosis indicates a peaked data distribution in comparison to the normal distribution that has zero kurtosis. The skewness is a measure of the symmetry of the data distribution. Negative skewness indicates a bias to the left, and positive skewness indicates a bias to the right.

Table 1: Data sets statistical summary. Mean, standard deviation, minimum, maximum in thousands of €.

Id	Length	Null %	Mean	Std	Kurtosis	Skewness	Min	Max
1	856	35,7	0,01	3,38	594,81	22,37	-9,07	90,27
2	684	29,8	0,26	5,80	58,98	3,69	-56,51	62,66
3	856	8,5	0,36	35,35	163,62	6,28	-303,20	671,04
4	1201	34,9	-0,12	14,32	78,14	-6,30	-223,38	72,76
5	849	19,4	0,00	1,67	56,10	-0,48	-18,26	16,42
6	799	20,7	0,01	6,63	33,21	-2,42	-68,97	56,27
7	772	38,5	0,07	5,36	86,75	6,74	-24,41	82,91
8	695	21,7	0,05	3,15	14,27	-2,57	-24,21	11,31
9	852	18,8	0,73	56,54	18,92	-0,78	-411,41	473,36
10	744	13,2	0,12	6,95	70,63	0,60	-81,13	78,72
11	639	62,6	-0,05	8,56	391,86	-17,65	-191,53	30,74
12	503	2,6	0,48	35,30	449,38	20,70	-47,27	771,38
13	697	24,7	0,52	24,24	18,81	2,06	-99,39	227,45
14	604	4,6	0,10	13,23	8,51	1,05	-63,23	92,71
15	605	4,1	0,68	11,67	4,43	0,33	-54,75	55,61
16	596	6,4	0,01	1,46	107,82	6,68	-8,48	22,61
17	1102	25,1	0,58	13,31	215,97	11,96	-118,01	250,13
18	552	3,1	0,16	2,16	70,23	5,10	-16,14	26,36
19	503	2,4	-0,31	2,58	6,43	0,50	-15,06	15,28
20	848	27,8	0,02	1,07	96,19	3,86	-12,07	16,04
21	829	18,7	-0,06	5,99	33,36	-1,62	-70,00	53,17
22	494	1,6	-0,46	27,28	22,64	-1,96	-244,29	138,87
23	604	9,1	1,63	20,85	79,99	5,41	-124,19	269,27
24	1097	8,4	0,96	20,36	95,45	6,48	-73,33	317,85
25	587	10,9	0,49	13,94	119,60	6,93	-116,01	201,13
26	751	11,6	-0,02	1,77	15,73	0,15	-10,73	15,56
27	332	8,1	0,29	1,64	10,60	2,14	-4,36	11,84
28	855	5,1	0,00	4,64	13,83	1,77	-18,10	39,01
29	609	13,6	0,04	6,07	108,66	-6,35	-90,04	55,89
30	554	8,1	0,03	1,47	68,26	5,47	-4,81	19,82
31	372	29,6	0,37	8,05	31,46	-2,41	-80,44	34,95
32	1103	24,8	0,28	4,03	11,07	0,54	-25,76	24,50
33	854	31,0	-0,19	6,81	115,63	-1,74	-94,33	95,59
34	1508	11,5	-0,06	10,13	19,89	-2,32	-96,82	49,65
35	501	7,4	0,20	5,40	11,41	-0,58	-31,42	29,19
36	359	11,4	0,42	1,85	12,24	2,44	-7,87	11,84
37	361	3,0	-0,69	17,82	139,06	-1,38	-228,88	218,42
38	170	9,4	-1,20	7,10	43,34	-5,73	-61,93	19,66
39	1104	29,0	0,02	0,95	7,95	-0,07	-5,67	6,57
40	198	0,0	0,78	12,38	0,58	1,02	-25,63	36,91
41	341	17,6	-0,25	8,34	15,80	1,22	-44,29	64,34
42	566	11,0	0,01	1,82	308,62	-15,80	-37,02	7,48
43	750	3,2	0,34	13,10	7,66	-0,04	-65,84	73,40
44	287	4,2	0,52	11,46	81,19	-0,05	-118,74	120,34
45	1465	49,8	0,04	9,12	43,51	-2,89	-107,20	75,47
46	565	44,8	0,54	5,58	75,41	2,91	-51,16	73,83
47	503	4,4	1,98	46,81	46,03	1,37	-338,39	478,26
48	605	13,1	0,21	22,71	34,31	-1,68	-207,04	203,09
49	993	50,5	-0,08	1,36	27,18	-2,18	-10,78	12,73
50	605	45,0	-0,01	27,37	43,79	-2,01	-262,52	221,96
51	1225	0,2	15,09	96,96	2,77	0,12	-419,88	481,66
52	1225	0,4	8,94	49,39	36,23	2,81	-325,46	700,66
53	1223	39,7	0,47	9,13	203,12	-10,25	-196,88	38,48
54	1225	52,3	0,46	77,91	151,93	4,28	-1021,36	1532,10

2.1 Normality

Next, we study if our cash flows follow a Gaussian distribution. The observed kurtosis and skewness can be used as a first normality test of the data distribution for each company. Table 1 shows that no company presents zero kurtosis and skewness. Only company 40, with kurtosis 0.58 and skewness 1.02, could be considered close to normality. Two additional tests can be used to either verify or reject the hypothesis of normality: the Shapiro-Wilk test for normality (Royston, 1982) and the Lilliefors (Kolmogorov-Smirnov) test for normality (Lilliefors, 1967). The results from these two tests allow us to reject the hypothesis of normally distributed cash flows for all the companies in our data set.

As pointed out elsewhere (Emery, 1981; Pindado and Vico, 1996), a possible explanation for non-normality could be the presence of abnormally high values or heavy tails. Thus, we repeated the Shapiro-Wilk test and the Lilliefors (Kolmogorov-Simirnov) test for normality, but using a trimmed version of the net cash flow time series by deleting observations greater or lower than three times the sample standard deviation. No difference in the results of the tests is observed, confirming the non-normality hypothesis beyond the conjectures of Emery (1981) and Pindado and Vico (1996).

Non-normal residuals may be problematic in the estimation process when using linear models. Data transformations such as the Box and Cox (1964) transformation to normality represent a possible solution. Forecasts are then calculated on the transformed data, but we must reverse the transformation to obtain forecasts on the original data, resulting in two additional steps. However, these transformations are not always the solution to the non-normality problem. Using both the original observations and the trimmed version of our data sets, we proceeded to transform the data using a Box-Cox transformation of the type:

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if} \quad \lambda_1 \neq 0, \\ \log(y+\lambda_2) & \text{if} \quad \lambda_1 = 0, \end{cases} \tag{1}$$

where y is the original time series, and λ_1 and λ_2 are parameters. In these experiments, we obtained parameter λ_1 by maximizing the log-likelihood of a linear estimation of the time series based on day-of-month and day-of-week dummy variables. We set parameter λ_2 to minus two times the minimum value of the time series. Then, we repeated the Lilliefors (Kolmogorov-Smirnov) test for normality obtaining again negative results for normality, both for the original and for the trimmed version of our data sets. As a result, even after Box-Cox transformation, the normality hypothesis does not hold.

2.2 Independence and seasonality

In what follows, we test the independence of cash flows and we also explore if seasonality is present. Autoregressive Integrated Moving Average (ARIMA) models by Box and Jenkins (1976), have been extensively used for time series analysis and forecasting. When dealing with time series, the autocorrelation coefficient, r_k, describes the relationship between observations that are lagged k time periods (Makridakis et al., 2008). We say that a time series is independent when the r_k values for different lags are close to zero. An example of an independent time series is the so-called white-noise model where each observation is made by adding a random component to a certain level.

An intuitive plot to assess independence is the Poincaré map (Kantz and Schreiber, 2004), which is a scatter plot of the original time series and a k-periods lagged time series as in Figure 2, which shows a lag of 1 day for time series 1 and 2 from Table 1. As a reference, we also include the Poincaré map for a Gaussian noise and for a sinusoidal time series. A cloud of points suggests independence, as for time series 1 and Gaussian noise, and the presence of any form suggests a more complex relationship, as for time series 2 and the sinusoidal.

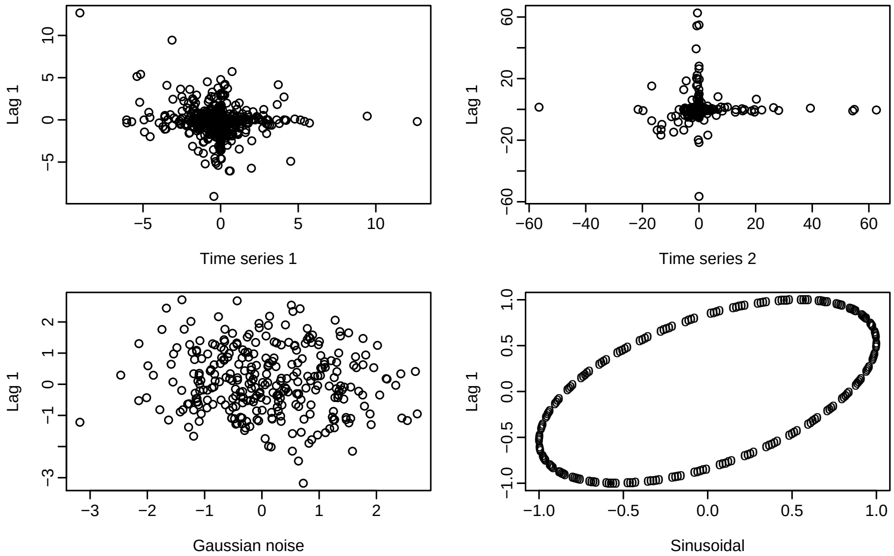


Figure 2: Poincaré map with lag 1 for time series 1 and 2.

A more general approach is to consider a set of the first r_k values as a whole as in the Ljung and Box (1978) test, which we applied to our data and produced mixed results. More precisely, we found that the hypothesis of independence could not be rejected in 24 out of 54 companies as summarized in Table 2. On the one hand, these results imply an underlying stochastic process for those data sets presenting independence. On the other hand, some kind of serial correlation such as seasonality is likely to be present in the case of companies presenting a certain degree of dependence in the sample. Seasonality is defined as the existence of a pattern that repeats itself over fixed time intervals in the data (Makridakis et al., 2008). It can be identified by significant autocorrelation coefficients. Seasonal trend decomposition methods (Cleveland et al., 1990), seasonal ARIMA models (Box and Jenkins, 1976; Franses and Van Dijk, 2005) or linear (and non-linear) regression models based on seasonal variables are available options to deal with seasonality. In cash flow forecasting, the distribution approach by Miller and Stone (1985) also deserves to be mentioned.

As mentioned in the introduction, previous works by Emery (1981); Miller and Stone (1985); Stone and Miller (1987); and Pindado and Vico (1996), reported the influence of day-of-month and day-of-week effects on cash flow patterns. Here, we test the presence of seasonality by fitting a regression model on daily cash flows using day-of-month and day-of-week dummy variables. Then, we use the F-test to test the significance of the overall regression. Table 2 reports, on the one hand, the Ljung-Box independence test and, on the other hand, the F-statistic, the p-value and the coefficient of determination R^2, derived from the regression. Interestingly, not all the data sets whose independence tests were rejected implied a significant regression based on dummy variables. Non-linear patterns, non-periodical temporal correlations, and the effect of outliers become possible explanations as we will see below.

Table 2: Independence and seasonality test results.

Id	Ljung Box test	F-statistic	p-value	R^2
1	Non-rejected	1,99	<0,05	0,08
2	Rejected	1,05	0,39	0,05
3	Non-rejected	1,87	<0,05	0,07
4	Rejected	1,51	<0,05	0,04
5	Rejected	1,85	<0,05	0,07
6	Non-rejected	1,12	0,29	0,05
7	Rejected	5,47	<0,05	0,20
8	Rejected	0,79	0,80	0,04
9	Rejected	5,30	<0,05	0,18
10	Rejected	2,04	<0,05	0,09
11	Non-rejected	0,97	0,51	0,05
12	Non-rejected	0,98	0,51	0,07
13	Rejected	5,21	<0,05	0,21
14	Rejected	7,13	<0,05	0,30
15	Rejected	1,92	<0,05	0,10
16	Non-rejected	4,31	<0,05	0,21
17	Non-rejected	4,91	<0,05	0,14
18	Rejected	2,99	<0,05	0,16
19	Rejected	2,58	<0,05	0,16
20	Non-rejected	2,71	<0,05	0,10
21	Non-rejected	1,37	0,08	0,06
22	Non-rejected	1,49	<0,05	0,10
23	Rejected	5,60	<0,05	0,25
24	Non-rejected	15,41	<0,05	0,33
25	Non-rejected	4,23	<0,05	0,21
26	Rejected	1,22	0,18	0,05
27	Non-rejected	1,24	0,18	0,12
28	Rejected	5,64	<0,05	0,19
29	Non-rejected	1,37	0,08	0,08
30	Rejected	6,18	<0,05	0,29
31	Non-rejected	1,25	0,16	0,11
32	Rejected	4,81	<0,05	0,13
33	Rejected	1,57	<0,05	0,06
34	Rejected	11,61	<0,05	0,21
35	Rejected	0,99	0,49	0,07
36	Non-rejected	1,82	<0,05	0,16
37	Rejected	1,58	<0,05	0,14
38	Non-rejected	1,06	0,39	0,21
39	Rejected	6,11	<0,05	0,16
40	Rejected	0,86	0,68	0,15
41	Non-rejected	1,72	<0,05	0,16
42	Non-rejected	3,90	<0,05	0,20
43	Rejected	2,96	<0,05	0,12
44	Non-rejected	1,89	<0,05	0,20
45	Rejected	1,26	0,15	0,03
46	Non-rejected	1,32	0,11	0,08
47	Non-rejected	0,90	0,63	0,06
48	Non-rejected	1,71	<0,05	0,09
49	Rejected	26,15	<0,05	0,48
50	Rejected	1,24	0,17	0,07
51	Rejected	16,66	<0,05	0,32
52	Rejected	5,01	<0,05	0,13
53	Non-rejected	1,59	<0,05	0,04
54	Rejected	0,88	0,67	0,02

2.3 Stationarity

In this section, we analyze if cash flows from our data set can be labeled as stationary. Basically, stationarity means that there is no drift in the time series behavior over time. We can visually assess stationarity by inspecting a time series plot as the one shown in Figure 1. Virtually, every process we find in nature is non-stationary, since its parameters depend on time (Kantz and Schreiber, 2004). However, a minimum requirement is that basic statistical properties of a distribution, such as mean and variance, remain constant over time, when measured through appropriately long time windows.

Emery (1981) studied stationarity by applying the Kolmogorov-Smirnov test for normality of cash flow by months. For comparative purposes, we applied the same procedure and, if any of the monthly tests rejected the hypothesis of normality, the whole time series was considered non-stationary. Following this procedure, only company 43 could be considered stationary.

Following the recommendations in Kantz and Schreiber (2004), we also perform a quick-and-dirty stationarity test from the fluctuations of a sample mean and variance. More precisely, we compute the sample mean and variance of each time series by months and obtain the standard errors for both. If the observed fluctuations of the running mean and variance are within these errors, then we consider the time series stationary. The results from this test shows that none of the time series is stationary either in mean or variance.

One way of removing non-stationarity is time series differencing, which can be defined as the change between two consecutive observations. After differencing, we repeated our simple test obtaining slightly different results. Only 10 out of 54 time series can be considered stationary in mean but none of them can be considered stationary in variance, a phenomenon known as heteroskedasticity. Thus, as a result, we conclude that our cash flow time series are non-stationary in mean and variance, even after differencing.

2.4 Discussion

Our results show that the widely extended hypothesis of cash flow normality is not present in our data sets. The presence of high abnormal values does not explain this behavior since non-normality persisted after removing these abnormal values. Non-linearity could be a possible explanation as we will see below. We also reported mixed results on independence and the influence of day-of-month and day-of-week effects on cash flow along the lines of the literature. We additionally report that common solutions to non-normality and non-stationarity such as data transformation and differencing produced little benefit when applied to our data set. Since seasonality and serial correlation are also present in our data set, we further explore the utility of alternative forecasting models. More precisely, we next study linearity and data transformation as an additional part of our empirical analysis for cash flow forecasting.

3 A simple cross-validated test for non-linearity

Most forecasting models are linear for computational convenience. However, non-linear patterns are likely to be present in finance and business time series. A time series linear model is defined as a variable y_t that depends on an explanatory vector $\boldsymbol{x_t}$ for any time t as follows:

$$y_t = \boldsymbol{\beta' x_t} + e_t \tag{2}$$

where $\boldsymbol{\beta'}$ is a transposed vector of coefficients, and e_t is the error or the residual component. An alternative and more general model can also be considered:

$$y_t = g(\boldsymbol{x_t}) + \epsilon_t. \tag{3}$$

Different tests of linearity can be found in Ramsey (1969); Keenan (1985); Granger et al. (1993); Lee et al. (1993), and Castle and Hendry (2010). Basically, all of them follow a common approach: first, they choose a regression equation $g(\boldsymbol{x_t})$ in (3) including linear and non-linear terms and, second, they test for the significance of the non-linear terms. However, these approaches do not accommodate well for forecasting purposes due to the following reasons: (i) the assumption of a specific form $g(\boldsymbol{x_t})$ for the regression equation such as quadratic, cubic or exponential forms; (ii) cross-validation is neglected.

If we relax the assumption of the regression equation, different non-linear models such as random forests (Breiman, 2001), neural networks (Hornik et al., 1989; Zhang et al., 1998), or radial basis functions (Broomhead and Lowe, 1988), could also be considered. However, the consideration of non-linear functions may lead to overfitting to the original time series. To prevent this problem, we propose the use of time series cross-validation. Cross-validation is a method to assess the predictive performance of a forecasting model that circumvents the problem of overfitting the data by testing the accuracy of the model on subset of data not used in the estimation (Hyndman and Athanasopoulos, 2013). As a result, we here propose a simple cross-validated test for non-linearity based on the following steps:

1. Estimate two alternative forecasting models, one linear and another one non-linear.

2. Cross-validate the predictive accuracy of both models with respect to a baseline.

3. Label as trivial if both models are significantly worse than the baseline.

4. Label as non-linear if the error of the non-linear model is significantly lower than that of the linear model. Otherwise, label as linear.

We argue that our ultimate goal is to discover the model with higher accuracy, disregarding if it is linear or non-linear. Thus, we define non-linearity as a model dependent property. More precisely, we label a data set as non-linear if the normalized squared error of the non-linear model is significantly lower than that achieved using the linear model as described in Figure 3. Normalization is achieved through the comparison to a baseline. For statistically significant differences in performance between models, we use a Wilcoxon signed rank test of the null hypothesis that the distribution of the difference is symmetric about zero with a 95% confidence interval (Wilcoxon et al., 1970).

We mentioned in Section 2 that a common practice to assess the utility of forecasts derived from any model is to compare its accuracy to that of a baseline forecasting model. We here use a mean forecaster, meaning that forecasts are always the average of past observations. The use of a baseline allows us to label our data sets as trivial if neither the linear model nor the non-linear model are able to improve the accuracy of the baseline.

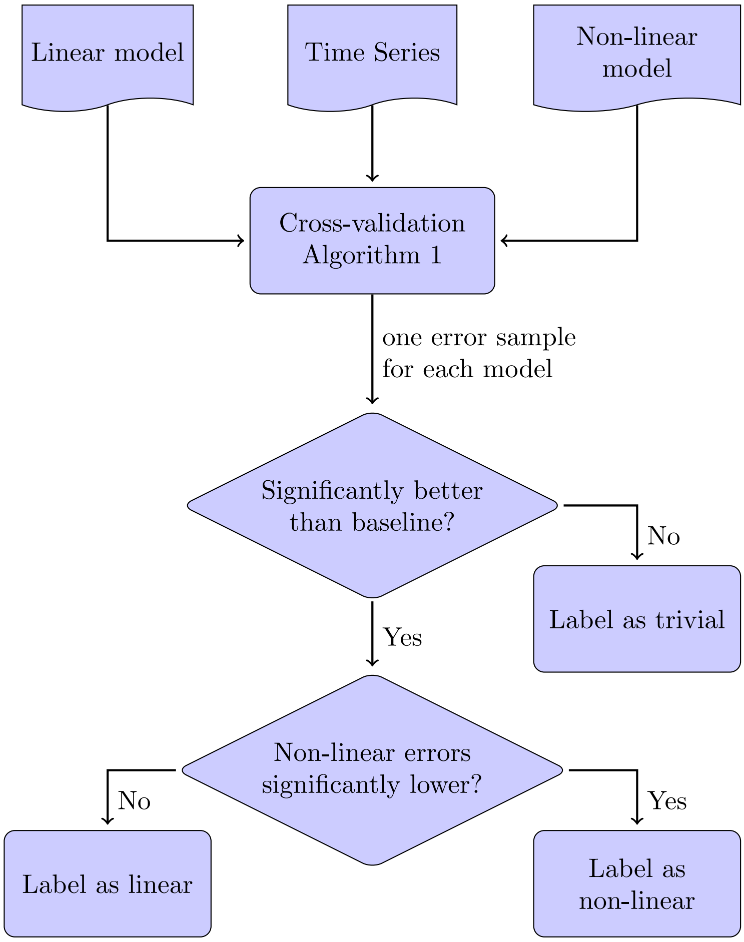


Figure 3: Simplified flow chart for our cross-validated test for non-linearity.

As detailed in Algorithm 1, we consider the minimum length k to estimate a model as the 80% of the oldest instances forming the training set. The remaining 20% of the instances form the test set for cross-validation. Initially, both the linear and the non-linear model are estimated using the first 80% of the instances. Then, forecasts for a prediction horizon up to 20 days are computed using the estimated models and squared errors are recorded. The estimation and forecasting process is repeated for each remaining instance in the test set, resulting into two paired error samples.

Algorithm 1: Algorithm for a simple cross-validation test for non-linearity

Input: Cash flow data set of T instances, minimum number k of instances to estimate a model, baseline m_0, linear model m_1, non-linear model m_2, prediction horizon h, level of significance α;
Output: Average prediction error for different prediction horizons, statistic for the difference in means, confidence interval;
for $i = 1, 2, \ldots, T - k - h + 1$ **do**
    Select the instances from time $k + i$ to $k + h + i - 1$, for the test set;
    Estimate m_0 with instances at times $1, 2, \ldots, k + i - 1$;
    Estimate m_1 with instances at times $1, 2, \ldots, k + i - 1$;
    Estimate m_2 with instances at times $1, 2, \ldots, k + i - 1$;
    Compute test errors ε_0, ε_1, ε_2 from time $k + i$ to $k + h + i - 1$;
end
Compute average h-step errors $\varepsilon_0(h)$, $\varepsilon_1(h)$, $\varepsilon_2(h)$;
Test for α significant differences between $\varepsilon_0(h)$, $\varepsilon_1(h)$, $\varepsilon_2(h)$;
if $\varepsilon_0(h) < \varepsilon_1(h)$ ***and*** $\varepsilon_0(h) < \varepsilon_2(h)$ **then**
    Label as trivial;
    else if $\varepsilon_2(h) < \varepsilon_1(h)$ **then**
        Label as non-linear;
        else
            Label as linear;
        end
    end
end

Although recent examples can be found in Kane et al. (2014); Mei et al. (2014); Zagorecki (2015), time series forecasting using random forests is not as well established as other non-linear models (Kantz and Schreiber, 2004; Clements et al., 2004; De Gooijer and Hyndman, 2006). Because of that and for illustrative purposes, we here consider a linear regression model and a non-linear random forest model, both using day-of-month and day-of-week variables as predictors. In the case of the linear regression model, each instance contains 34 dummy predictor variables, 30 for day-of-month and 4 for day-of week, and a cash flow observation. In the case of the random forest model, each instance contains 2 categorical variables, one for day-of-month and one for day-of-week.

Our results, summarized in Table 3, show that only about half of the data sets can be labeled as non-trivial because neither the linear model nor the non-linear model were able to significantly beat the trivial forecaster. From those data sets in which the hypothesis of independence could not be rejected (see Table 2), 20 out of 24 were labeled as trivial, confirming a stochastic behavior. On the other hand, only 6 of them were labeled as non-linear according to our cross-validated definition. However, it is important to note that our cross-validated definition of non-linearity depends on the alternative models considered and, consequently, on the predictive ability of the predictors used. If the time series is not a pure random process, then the search for a more informative set of features is meant to play a key role.

One may assume either linearity or non-linearity from the results of our non-

Table 3: Results of the test for non-linearity. Reg NSE = Regression normalized squared error; RF NSE = Random Forest normalized squared error

Id	Reg NSE	RF NSE	Statistic	p-value	Triviality	Linearity
1	0,99	1,00	26	<0,05	Non-Trivial	Linear
3	0,99	1,01	8	<0,05	Non-Trivial	Linear
4	1,00	1,01	0	<0,05	Non-Trivial	Linear
7	0,81	0,83	0	<0,05	Non-Trivial	Linear
9	0,90	0,93	3	<0,05	Non-Trivial	Linear
13	0,86	0,88	13	<0,05	Non-Trivial	Linear
14	0,76	0,77	45	<0,05	Non-Trivial	Linear
16	0,85	0,86	64	0,13	Non-Trivial	Linear
18	0,86	0,88	63	0,12	Non-Trivial	Linear
19	0,96	0,94	182	<0,05	Non-Trivial	Non-linear
20	0,99	0,98	209	<0,05	Non-Trivial	Non-linear
23	0,78	0,79	78	0,33	Non-Trivial	Linear
24	0,73	0,79	0	<0,05	Non-Trivial	Linear
25	0,77	0,81	21	<0,05	Non-Trivial	Linear
28	0,84	0,90	0	<0,05	Non-Trivial	Linear
29	0,99	0,99	30	<0,05	Non-Trivial	Linear
30	0,73	0,80	5	<0,05	Non-Trivial	Linear
33	0,94	0,93	166	<0,05	Non-Trivial	Non-linear
34	0,97	0,95	172	<0,05	Non-Trivial	Non-linear
39	0,96	0,96	36	<0,05	Non-Trivial	Linear
42	0,88	0,87	149	0,11	Non-Trivial	Linear
43	0,99	0,96	210	<0,05	Non-Trivial	Non-linear
48	1,01	0,99	191	<0,05	Non-Trivial	Non-linear
49	0,63	0,65	7	<0,05	Non-Trivial	Linear
51	0,77	0,80	0	<0,05	Non-Trivial	Linear
52	0,94	0,94	116	0,70	Non-Trivial	Linear

linearity test, but it is important to analyze the robustness of these results to both the presence of outliers and the impact of other data transformations.

4 The impact of data transformations

In this section, we aim to analyze the impact of outlier treatments on noise reduction, as intended, and on information loss, as an undesirable effect. We also study the influence of Box-Cox data transformations on the results of our cross-validated non-linearity test. Detection and treatment of outliers is an ongoing issue in data mining (Rousseeuw and Leroy, 1987; Hodge and Austin, 2004). An outlier is an observation that appears to significantly deviate from other members of the sample in which it occurs (Grubbs, 1969). Outliers arise due to changes in systems, measurement errors or simply due to deviations in data. It is also important to note that an outlier may also be the most interesting part of the data.

On the one hand, from the set of cash flow time series labeled as trivial, some of them may be labeled as non-trivial after removing outliers as a way of noise reduction. On the other hand, from those data sets labeled as non-trivial, some of them may be labeled as trivial due to the information loss produced by the treatment. We here measure the effect of removing outliers on the prediction error using time series cross validation for different thresholds of outlier replacement. For each data set, we progressively identify as outliers cash flow observations greater than 5, 4, and 3 times the standard deviation in a training set with the 80% oldest observations. We replace outliers with a linear interpolation and proceed as detailed in Algorithm 1 to cross-validate triviality and linearity. The results from this analysis are summarized in Table 4.

By following this procedure, we identified data sets 5, 10, 17, 32, 44 and 54 (6 out of 28), initially labeled as trivial that, after outlier treatment, could be labeled as non-trivial due to noise reduction. Similarly, data sets 4 and 48 that were initially labeled as non-trivial could be labeled as trivial after outlier treatment due to information loss. If we measure noise reduction by the error reduction and information loss by the error increase, then we can

Table 4: Results of the test for non-linearity after outlier treatment and Box-Cox transformation. Changes in labels are marked with *.

		After outliers		After outliers and Box-Cox	
Id	Triviality	Linearity	Noise reduction	Linearity	Noise reduction
1	Non-Trivial	Linear	0,00	Non-linear*	-0,01
3	Non-Trivial	Linear	0,02	Non-linear*	0,00
5	Non-Trivial	Non-linear	0,40	Non-linear	0,41
7	Non-Trivial	Linear	-0,10	Linear	-0,13
9	Non-Trivial	Linear	-0,04	Linear	-0,04
10	Non-Trivial	Non-linear	0,46	Non-linear	0,47
13	Non-Trivial	Linear	-0,18	Linear	-0,21
14	Non-Trivial	Linear	-0,05	Linear	-0,07
16	Non-Trivial	Linear	-0,18	Linear	-0,17
17	Non-Trivial	Non-linear*	0,71	Non-linear	0,71
18	Non-Trivial	Non-linear*	-0,20	Non-linear	-0,20
19	Non-Trivial	Non-linear	-0,03	Non-linear	-0,04
20	Non-Trivial	Non-linear	-0,02	Non-linear	-0,02
23	Non-Trivial	Non-linear*	-0,22	Non-linear	-0,22
24	Non-Trivial	Linear	-0,20	Linear	-0,06
25	Non-Trivial	Non-linear*	-0,26	Non-linear	-0,25
28	Non-Trivial	Linear	-0,05	Linear	-0,04
29	Non-Trivial	Linear	0,07	Non-linear*	0,00
30	Non-Trivial	Linear	-0,06	Linear	-0,04
32	Non-Trivial	Non-linear	0,18	Non-linear	0,21
33	Non-Trivial	Linear*	-0,12	Linear	-0,11
34	Non-Trivial	Linear*	0,12	Linear	0,09
39	Non-Trivial	Non-linear*	-0,02	Linear*	-0,01
42	Non-Trivial	Linear	-0,23	Linear	-0,14
43	Non-Trivial	Non-linear	0,04	Non-linear	0,03
44	Non-Trivial	Non-linear*	0,48	Non-linear	0,82
49	Non-Trivial	Non-linear*	-0,56	Non-linear	-0,61
51	Non-Trivial	Linear	-0,03	Linear	-0,03
52	Non-Trivial	Linear	0,01	Linear	0,03
54	Non-Trivial	Linear*	0,17	Linear	0,17

assess the impact of outlier treatment. Following this approach, we obtained mixed results for non-trivial data sets after outlier treatment: an average noise reduction of 22%, and an average information loss of 14%. It is important to recall that unexpected observations are often the most interesting part of the data to predict. For instance, when the goal is to forecast unusual but genuine cash flows.

Non-linearity and outliers are closely linked. Indeed, Castle and Hendry (2012) hypothesized that non-linear functions can align with outliers, causing functions to be considered relevant spuriously, which can be detrimental for generalizing and forecasting. If this hypothesis is correct, the relative forecasting ability of a linear model in comparison to a non-linear model would increase as the presence of outliers in a training set is reduced. From the set of time series finally labeled as non-trivial, data sets 33, 34 and 54, initially labeled as non-linear changed their labels to linear. Surprisingly, data sets 17, 18, 23, 25, 39, 44 and 49 (7 out of 30), could be labeled as non-linear after outlier treatment. Except for data sets 17 and 44, in all cases there was information loss, i.e., error increase, suggesting that non-linear models can deal better with information loss.

We also considered a Box-Cox transformation to analyze if this kind of data transformation may influence the results from our cross-validated non-linearity test. From the set of non-trivial data sets we compare linearity labels, first, after outlier treatment, and second, after outlier treatment and Box-Cox transformation as described in equation (1). In addition, we compare information loss computed as the difference between the sum of errors of the linear and non-linear forecasting models before and after the outlier treatment. A positive value means noise reduction or error reduction while a negative value means information loss or error increase. These results show that our cross-validated non-linearity test outputs similar results after Box-Cox transformation since the change in labels were produced in data sets with similar linear and non-linear model performance.

A summary of the results of this section is shown in Table 5. The high number of trivial data sets may be caused by the general inherent randomness of cash flows. Outlier treatment produced a small improvement in non-triviality but also an outstanding increase in non-linearity. Finally, Box-Cox data transformation yielded similar results but with better results for non-linear models. Thus, we conclude that: (i) common data transformations had little impact on our time series in terms of linearity; and (ii) outlier treatment and Box-Cox transformation were unable to transform non-linear into linear cash flows.

Table 5: Number of time series data sets and their labels after transformation. OT=Outlier treatment; DT=Data transformation.

Label	Raw data	After OT	After OT and DT
Trivial	28	24	24
Non-trivial	26	30	30
Linear	20	17	15
Non-linear	6	13	15

5 Concluding remarks

In this paper, we provide a complete empirical study of the statistical properties of daily cash flows based on 54 real-world cash flow time series from small and medium companies. This study is the most comprehensive empirical study on daily cash flow so far in terms of the range of statistical properties considered, and on both the number and the length of the data sets that we make available online. We focus on the implications for forecasting due to its key role in cash management.

5.1 Summary of findings

Our results show that the extended hypothesis of normal, stationary and independent cash flows is hardly present in our cash flow data set. Thus, we conclude that assumptions of normality, stationarity and independence that have been extensively used in cash management literature are not realistic. We also highlight that common solutions to non-normality and non-stationarity such as data transformation and differencing produce little benefit when applied to our data sets, with the risk of loosing important information on extreme cash flows.

In an attempt to discover the attributes of actual-world cash flows, we also studied the presence of non-linearity. To this end, we proposed a new simple test for non-linearity with two main advantages in comparison to alternative approaches. First, our test does not assume any non-linear function. Second, it is based on time series cross validation to increase robustness and avoid overfitting. It is important to note that our cross-validated definition of non-linearity depends on the alternative models considered, one linear and another one non-linear.

Our simple cross-validated non-linearity test labeled as either trivial, linear or non-linear our cash flow data set after outlier treatment resulting in an important increase the number of data sets labeled as non-linear. After both outlier treatment and Box-Cox transformation, linearity could not be achieved and non-linear models showed more robust.

5.2 Implications

Our results raise questions about two common assumptions in cash flow time series since we found that: (i) the usual assumption of normality, independence and stationarity is hardly present; and (ii) common data transformations such as outlier treatment and Box-Cox transformation have little impact on normality and linearity. Contrary to the rather common assumption in the literature, these results imply that neither it is always possible to achieve a Gaussian, noise-free and linear time series through data transformation nor it is always desirable due to information loss. Thus, linear models should be considered as an initial step towards more realistic ones which are better adapted to real cash flow situations. The results from our cross-validated test for non-linearity suggest that non-linear models represent a justifiable alternative for time series forecasting. Moreover, since our test is both model and outlier dependent, a promising line of future work would be the integration of outlier treatment in the test itself in an attempt to assess noise reduction or information loss.

As a result, we claim that a number of preliminary steps are necessary in cash flow forecasting before model selection: (i) statistical summary including normality, independence and stationarity; (ii) impact of data transformations such as outlier treatment and Box-Cox transformation; (iii) non-linearity test to determine the type of model which is expected to deliver a better performance. Finally, this process is not limited to daily cash flow, since it can also be applied to any other time series data set when cross-validation is required.

References

References

Baccarin, S. (2009). Optimal impulse control for a multidimensional cash management system with generalized cost functions. *European Journal of Operational Research*, 196(1):198–206.

Baumol, W. J. (1952). The transactions demand for cash: An inventory theoretic approach. *The Quarterly Journal of Economics*, 66(4):545–556.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Box, G. E. and Jenkins, G. M. (1976). *Time series analysis: forecasting and control, revised ed.* Holden-Day.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Broomhead, D. and Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2(3):321–355.

Castle, J. L. and Hendry, D. F. (2010). A low-dimension portmanteau test for non-linearity. *Journal of Econometrics*, 158(2):231–245.

Castle, J. L. and Hendry, D. F. (2012). Automatic selection for non-linear models. In *System Identification, Environmental Modelling, and Control System Design*, pages 229–250. Springer.

Clements, M. P., Franses, P. H., and Swanson, N. R. (2004). Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, 20(2):169–183.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73.

Constantinides, G. M. and Richard, S. F. (1978). Existence of optimal simple policies for discounted-cost inventory and cash management in continuous time. *Operations Research*, 26(4):620–636.

da Costa Moraes, M. B., Nagano, M. S., and Sobreiro, V. A. (2015). Stochastic cash flow management models: A literature review since the 1980s. In *Decision Models in Engineering and Management*, pages 11–28. Springer International Publishing.

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473.

Emery, G. W. (1981). Some empirical evidence on the properties of daily cash flow. *Financial management*, 10(1):21–28.

Franses, P. H. and Van Dijk, D. (2005). The forecasting performance of various models for seasonality and nonlinearity for quarterly industrial production. *International Journal of Forecasting*, 21(1):87–102.

Gormley, F. M. and Meade, N. (2007). The utility of cash flow forecasts in the management of corporate cash balances. *European journal of operational research*, 182(2):923–935.

Granger, C. W., Terasvirta, T., et al. (1993). Modelling non-linear economic relationships. *OUP Catalogue*.

Gregory, G. (1976). Cash flow models: a review. *Omega*, 4(6):643–656.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126.

Homonoff, R. and Mullins, D. W. (1975). *Cash management: an inventory control limit approach*. Lexington Books.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hyndman, R. and Athanasopoulos, G. (2013). Forecasting: principles and practice. `http://otexts.org/fpp/`. Last accessed: 2015-05-01.

Kane, M. J., Price, N., Scotch, M., and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, 15(1):1.

Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis*, volume 7. Cambridge University Press.

Keenan, D. M. (1985). A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, 72(1):39–44.

Lee, T.-H., White, H., and Granger, C. W. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, 56(3):269–290.

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (2008). *Forecasting methods and applications*. John Wiley & Sons.

Mei, J., He, D., Harley, R., Habetler, T., and Qu, G. (2014). A random forest method for real-time price forecasting in new york electricity market. In *PES General Meeting— Conference & Exposition, 2014 IEEE*, pages 1–5. IEEE.

Miller, M. H. and Orr, D. (1966). A model of the demand for money by firms. *The Quarterly journal of economics*, 80(3):413–435.

Miller, T. W. and Stone, B. K. (1985). Daily cash forecasting and seasonal resolution: Alternative models and techniques for using the distribution approach. *Journal of Financial and Quantitative Analysis*, 20(03):335–351.

Penttinen, M. J. (1991). Myopic and stationary solutions for stochastic cash balance problems. *European journal of operational research*, 52(2):155–166.

Pindado, J. and Vico, J. (1996). Evidencia empírica sobre los flujos de caja. un nuevo enfoque en su tratamiento. *Revista Española de Financiación y Contabilidad*, 25(87):497–517.

Premachandra, I. (2004). A diffusion approximation model for managing cash in firms: An alternative approach to the miller–orr model. *European Journal of Operational Research*, 157(1):218–226.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(2):350–371.

Ross, S. A., Westerfield, R., and Jordan, B. D. (2002). *Fundamentals of corporate finance*. McGraw-Hill, sixth edition.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*, volume 589. John Wiley & Sons.

Royston, J. (1982). An extension of shapiro and wilk's w test for normality to large samples. *Applied Statistics*, 31(2):115–124.

Srinivasan, V. and Kim, Y. H. (1986). Deterministic cash flow management: state of the art and research directions. *Omega*, 14(2):145–166.

Stone, B. K. and Miller, T. W. (1981). Daily cash forecasting: a structuring framework. *Journal of Cash Management*, 1:35–50.

Stone, B. K. and Miller, T. W. (1987). Daily cash forecasting with multiplicative models of cash flow patterns. *Financial Management*, 16(4):45–54.

Stone, B. K. and Wood, R. A. (1977). Daily cash forecasting: a simple method for implementing the distribution approach. *Financial Management*, pages 40–50.

Wilcoxon, F., Katti, S., and Wilcox, R. A. (1970). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.

Zagorecki, A. (2015). Prediction of methane outbreaks in coal mines from multivariate time series using random forest. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pages 494–500. Springer.

Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62.