Tasmanian School of Business and Economics
University of Tasmania

Discussion Paper Series N 2021-06

A simple linear alternative to Multiplicative Error Models with an application to trading volume

Adam Clements

Queensland University of Technology, Australia

Stan Hurn

Queensland University of Technology, Australia

Vladimir Volkov

University of Tasmania, Australia

ISBN 978-1-922708-08-3

A simple linear alternative to Multiplicative Error Models with an application to trading volume*

Adam Clements[1], Stan Hurn[1], and Vladimir Volkov[2]

[1]School of Economics and Finance, Queensland University of Technology, Australia
[2]Tasmanian School of Business and Economics, University of Tasmania, Australia

December 2, 2020

Abstract

Forecasting intraday trading volume is an important problem in economics and finance. One influential approach to achieving this objective is the non-linear Component Multiplicative Error Model (CMEM) that captures time series dependence and intraday periodicity in volume. While the model is well suited to dealing with a non-negative time series, it is relatively cumbersome to implement. This paper proposes a system of linear equations, that is estimated using ordinary least squares, and provides at least as good a forecasting performance as that of the CMEM. This linear specification can easily be applied to model any time series that exhibits diurnal behaviour.

Keywords
Volume, forecasting, high-frequency data, CMEM, diurnal.

JEL Classification Numbers
C22, G00.

*Corresponding author: Vladimir Volkov - vladimir.volkov@utas.edu.au.

1 Introduction

The Volume Weighted Average Price (VWAP) trading strategy (Berkowitz et al., 1988, Madhavan, 2002, Bialkowski et al., 2008) is a popular strategy for improving trade execution, one that requires accurate predictions of intraday volume. Brownlees et al. (2011) propose a Component Multiplicative Error Model (CMEM) for modeling and forecasting intraday trading volume that is able to deal with both the persistence of and diurnal patterns in volume. They demonstrate that the CMEM performs well when forecasing out-of-sample.

This paper demonstrates that an alternative approach to forecasting intraday volume, based on simple system of linear equations, will perform as well, if not better, than the nonlinear CMEM approach. This multiple equation approach captures both the diurnal periodic component and an intraday dynamic component that are present in intra-day trading volume. The resultant system has more parameters than a standard CMEM model and cannot guarantee that the forecast will be positive, but it has the advantage that it can be estimated straightforwardly by ordinary least squares. An empirical application demonstrates the out-of-sample forecasts generated from the linear system are at least as good as the CMEM forecasts according to VWAP criterion, and better than the CMEM forecasts when judged on the MSE criterion.

2 The models

Let each trading day, $t \in \{1, ..., T\}$, be divided into equally spaced intervals (or bins), $i \in \{1, ..., I\}$ where in this application $I = 78$. Intraday trading volume on day t for bin i is then denoted $x_{t,i}$.

2.1 The CMEM model

The CMEM model (Brownlees et al., 2011) aims to model $x_{t,i}$ by means of three dynamic features, a daily component, an intra-day periodic component and an intra-day dynamic component. The three components of the CMEM are defined as follows.

(i) **Intraday periodic component:**

$$\phi_i = \exp\Big(\sum_{k=1}^{I/2}\Big[\gamma_k \cos\Big(\frac{2\pi}{I}k(i-1)\Big) + \delta_k \sin\Big(\frac{2\pi}{I}k(i-1)\Big)\Big], \tag{1}$$

with $I = 78$. Although the possible maximum number of terms in the expansion is $k = 36$, only the first 12 frequencies are used in this application. The intraday periodic component models the diurnality in intraday volume.

(ii) **Daily periodic component:**

$$\eta_t = \alpha_0 + \alpha_1 \eta_{t-1} + \alpha_2 x_{t-1}^{(\eta)} \tag{2}$$

where $x_t^{(\eta)}$ is the standardized daily volume

$$x_t^{(\eta)} = \frac{1}{I}\sum_{i=1}^{I} \frac{x_{t,i}}{\phi_i\,\mu_{t,i}}. \tag{3}$$

The daily component encapsulates the level of the series.

(iii) **Intraday dynamic component:**

$$\mu_{t,i} = \beta_0 + \beta_1 \mu_{t,i-1} + \beta_2 x_{t,i-1}^{(\mu)}, \tag{4}$$

where $x_{t,i}^{(\mu)}$ is the standardized intraday volume

$$x_{t,i}^{(\mu)} = \frac{x_{t,i}}{\phi_i\,\eta_t}. \tag{5}$$

The identification restriction, $\beta_0 = 1 - \beta_1 - \beta_2$ is imposed in the estimation. The intraday dynamic component takes care of any dynamic structure that is not periodic in nature.

The CMEM model is specified as

$$x_{t,i} = \eta_t\,\phi_i\,\mu_{t,i}\,\varepsilon_{t,i} \qquad \varepsilon_{t,i} \sim (1, \sigma^2). \tag{6}$$

The multiplicative error term, $\varepsilon_{t,i}$ is assumed to be nonnegative. The parameters of the CMEM are estimated by quasi-maximum likelihood estimation based on the assumption that the density of the disturbances follows a gamma distribution as in Engle and Gallo (2006). The log likelihood function at time t is given by

$$\log L_t = -\log\Gamma(a) + a\log(a) + (a-1)\log(x_{t,i}) - a\log(\eta_t\,\phi_i\,\mu_{t,i}) - a\Big(\frac{x_{t,i}}{\eta_t\,\phi_i\,\mu_{t,i}}\Big), \tag{7}$$

where a is the parameter of the gamma distribution. The estimation is performed in two steps. In the first step, the Fourier parameters for the seasonal factors ϕ_i are estimated by ordinary least squares regression. In the second step, conditional on the seasonal parameters obtained from the first step, the unknown parameters α_i, β_i and a are estimated.

2.2 The multiple equation approach

Multiple equation time series models have enjoyed some popularity in the literature but their influence has waned in recent years. Rather than trying to model the trajectory of time series of trading volume, each period (or bin) of the day is treated as a separate forecasting problem with its own equation (Peirson and Henley, 1994, Ramanathan et al., 1997, Espinoza et al., 2005, Soares and Medeiros, 2008). In other words, the data are re-arranged so that a daily time series is created for each of the 78 daily bins and a simple linear equation is used to model each of these 78 time series. The major advantage of this approach is that the model remains linear in parameters, so that ordinary least squares can be used to estimate the equations.

Each equation has a simple linear structure which involves the first and fifth daily (total volume for the day) lags of volume, x_{t-1} and x_{t-5}, respectively, and the volume and disturbance term from the immediately preceding bin, $x_{t,i-1}$ and $\widehat{v}_{t,i-1}$, respectively. The daily lags, x_{t-1} and x_{t-5} are included to capture longer-term persistence in volume (akin to the daily component in the CMEM) with the terms $x_{t,i-1}$ and $v_{t,i-1}$ used to capture intraday persistence. The first bin of the day is different to the others because there is no information from an immediately preceding bin on the same day. In this first bin, in place of $x_{t,i-1}$ and $v_{t,i-1}$, values from the final bin on the preceding day $t-1$ are used. This is in the same spirit as the intraday component in the CMEM, $\mu_{t,i}$ which links trading days together. The equation for the first bin $i = 1$ takes the form

$$x_{t,1} = \theta_{10} + \theta_{11}x_{t-1} + \theta_{12}x_{t-5} + \theta_{i3}x_{t-1,78} + \theta_{i4}v_{t-1,78} + v_{t,1}, \tag{8}$$

where θ_{10} a is constant term. For the remaining bins, $i = 2, \ldots, 78$, the linear equations are specified as

$$x_{t,i} = \theta_{i0} + \theta_{i1}x_{t-1} + \theta_{i2}x_{t-5} + \theta_{i3}x_{t,i-1} + \theta_{i4}v_{t,i-1} + v_{t,i}. \tag{9}$$

This structure offers a great deal in terms of flexibility. The set of intercepts deal with diurnal pattern, differing θ_{i1} and θ_{i2} coefficients allow for different impacts across the trading day, and θ_{i3} allows for varying persistence during the trading day. The structure is simple but works well for series with a strong diurnal structure.[1] The series of intercepts θ_{i0} pick up the diurnal pattern in volume and control the unconditional level for each bin. A daily lags of volume capture longer-term persistence present in daily trading volume. Where appropriate, intraday autoregressive persistence is captured with the term $x_{t,i-1}$ from the preceding bin, along with error from the previous bin, $\widehat{v}_{t,i-1}$.

The multiple equation model is estimated equation-by-equation using iterative ordinary least squares (Spliid, 1983). Each equation is initially estimated ignoring the $v_{t,i-1}$, $i = 2, \ldots, 78$ terms and the regression residuals stored. The equations are then re-estimated including $\widehat{v}_{t,i-1}$, $i = 2, \ldots, 78$ from the previous step as observed regressors. This process is then iterated until convergence which is defined as the difference in parameter values in successive iterations being less than a user supplied tolerance, in this case the square root of machine precision for floating-point arithmetic. While the CMEM approach guarantees that volume forecasts will be positive, this is not the case with multiple equation model. Even though theoretically speaking, negative forecasts are possible, at no point during any of the empirical analysis have negative forecasts been observed. In the context of volatility forecasting, the same issue arises. When the standard HAR model of Corsi (2009) is applied to a raw realized volatility series, there is no guarantee that the forecasts will be positive, however the risk of this occurring is negligible. Bollerslev et al. (2016) propose an extended version of the HAR model which does produce a very small number of negative forecasts. To mitigate this problem, they apply an 'insanity filter'

[1] A similar multi-equation model was used by Clements et al. (2016) to predict electricity load and by Moisan et al. (2018) to predict air-pollution concentrations and the linear approach is shown to outperform more complex nonlinear forecasting models. In both of these studies, more complex sets of explanatory variables are required than the current setting. Electricity load and air pollution exhibit a marked diurnal structure and also strong seasonality, with a range of different exogenous climatic and environmental variables needing to be included.

to their forecasts if their models produced forecasts outside of the observed range for volatility. A similar filter could be applied here should the situation demand it.

3 Empirical results

To compare forecast accuracy of CMEM and the mulitple equation linear system, the SPDR ETF (SPY) which tracks the S&P500 equity index is used. A dataset contains intraday volumes and transaction prices over a sample period from 5 January 2004 to 30 December 2016. The frequency of the dataset is 5 minutes (78 bins within a day). A detailed discussion of the empirical regularities of the SPY series is presented in Brownlees et al. (2011). The average daily volume of the SPY series (top panel) together with the familiar U-shaped intra-day pattern (bottom panel) are shown in Figure 1.

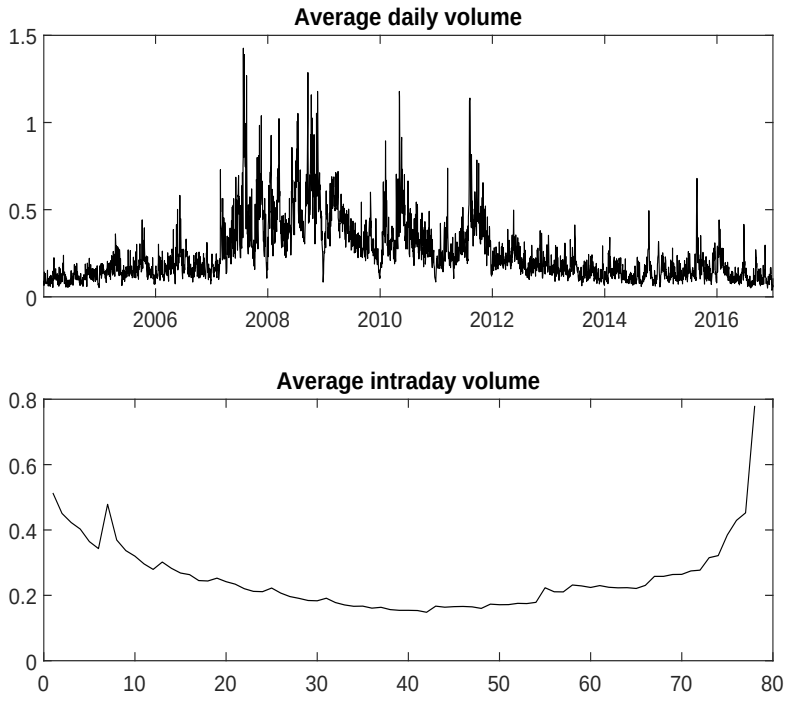


Figure 1: Average daily volume (top panel) and the diurnal pattern computed as the average volume in 5 minute intervals.

The estimated parameters of the CMEM are presented in Table 1. The results suggest that the level of daily volume captured by parameter α_0 is low and insignificant, a result that is confirmed by Naimoli and Storti (2019). Other parameters capturing daily dynamics, α_1 and α_2, are significant and close to those reported by Brownlees et al. (2011). Intraday parameter estimates show that this dynamic component $\mu_{t,i}$ follows an AR(1) structure as β_1 is close to 1 and β_2 is insignificant.

Detailed in-sample estimation results for the linear dynamic model from Section 2.2 are not

presented here but a brief outline of the salient features is as follows.The intercepts, $\hat{\theta}_{i0}$, are found to be largest at the beginning of trading day and then decrease during the middle of the day. This controls changes in the unconditional level reflecting the diurnal component. The Intraday autoregressive parameter estimates $\hat{\theta}_{i3}$ are lowest at the beginning of the trading day and gradually grow during the day. The reverse pattern is observed in the coefficients on the daily lag, $\hat{\theta}_{i1}$. The coefficients of the weekly lag, $\hat{\theta}_{i2}$ show that the effect of longer-term movements in volume are greatest as the start and end of the trading day. The coefficients on the lagged error, $\hat{\theta}_{i4}$ are negative and significant.

The autocorrelation of the estimated residuals of the linear model are reported in Figure 2. As noted by Naimoli and Storti (2019) the large number of intraday observations makes the autocorrelation tests extremely sensitive to deviations from the null hypothesis of white noise errors. Consequently, Figure 2 shows the ACFs for the first 50 days used for estimation. The residuals from the linear system exhibit minor negative correlations at the second and third lags. This pattern is very similar to the results presented in Figure 5 of Brownlees et al. (2011) for the basic model.

Table 1: Parameter estimates of a standard CMEM defined in Section 2.1. The sample period is from 5 Jan 2004 to 20 May 2011. The significant parameters at the 5% level are assigned (*).

		Parameter estimates
CMEM	β_1	0.98*
	β_2	0.01
	a	2.46*
	α_0	0.02
	α_1	0.58*
	α_2	0.40*

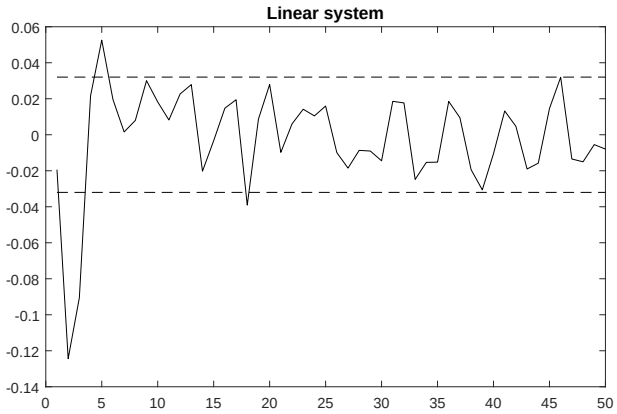


Figure 2: Autocorrelation of the residuals $v_{t,i}$ over the first 50 days. Confidence intervals are represented by $\pm 1.96 *$ standard errors.

Following Brownlees et al. (2011), two forecasting schemes are used here. One-day-ahead fore-

casts, $\widehat{x}_{t,i|t-1}$ are based on information available on the previous trading day and are made for all 78 intervals of the following day. The CMEM forecast is given by

$$\widehat{x}_{t,i|t-1} = \widehat{\eta}_{t|t-1}\,\widehat{\phi}_i\,\widehat{\mu}_{t,i|t-1}, \tag{10}$$

where $\widehat{\phi}_i$ signals that estimated coefficients are used in its construction and

$$\widehat{\eta}_{t|t-1} = \widehat{\alpha}_0 + \widehat{\alpha}_1\eta_{t-1} + \widehat{\alpha}_2 x^{(\eta)}_{t-1} \tag{11}$$

$$\widehat{\mu}_{t,i|t-1} = \widehat{\beta}_0 + \widehat{\beta}_1\widehat{\mu}_{t,i-1} + \widehat{\beta}_2\widehat{x}^{(\mu)}_{t,i-1|t-1}. \tag{12}$$

The recursion used for the multiple equation approach is again consists of forecasting for the first bin

$$\widehat{x}_{t,1|t-1} = \widehat{\theta}_{i0} + \widehat{\theta}_{i1}x_{t-1} + \widehat{\theta}_{i2}x_{t-5} + \widehat{\theta}_{i3}x_{t-1,78} + \widehat{\theta}_{i4}\widehat{v}_{t-1,78}, \tag{13}$$

and the remaining bins $i = 2, \ldots, 78$,

$$\widehat{x}_{t,i|t-1} = \widehat{\theta}_{i0} + \widehat{\theta}_{i1}x_{t-1} + \widehat{\theta}_{i2}x_{t-5} + \widehat{\theta}_{i3}\widehat{x}_{t,i-1|t-1} + \widehat{\theta}_{i4}\widehat{v}_{t,i-1|t-1}, \tag{14}$$

where $\widehat{v}_{t,i-1|t-1} = x_{t,i-1} - \widehat{x}_{t,i-1|t-1}$.

The second forecasting scheme is the one-bin-ahead forecast, $\widehat{x}_{t,i|i-1}$ that conditions on information from the preceding bins on the same trading day. The recursion only differs from that described above in that observed volume in the preceding bins are used in place of the forecasts. Therefore $x^{(\mu)}_{t,i-1}$ replaces $\widehat{x}^{(\mu)}_{t,i-1|t-1}$ in equation (12) and $x_{t,i-1}$ replaces $\widehat{x}_{t,i-1|t-1}$ in equation (14).

The out-of-sample forecasting procedure follows an expanding window approach. Both the system of equations and CMEM are initially estimated using a window size of 2000 days. As the out-of-sample forecasting procedure is conducted on a daily basis, all forecasts are made for all 78 intervals in the following day.

Initially, it is useful to compare average trading volume against average predicted intra-day volume over the whole sample (Figure 3). Interestingly both CMEM and the linear system capture the spike in trading activity after the first 5 bins, however high trading activity during the last 30 minutes of trading is captured better by the linear model.

To evaluate the out-of-sample forecast performance of CMEM and the linear model, following Brownlees et al. (2011), the following criteria are used:

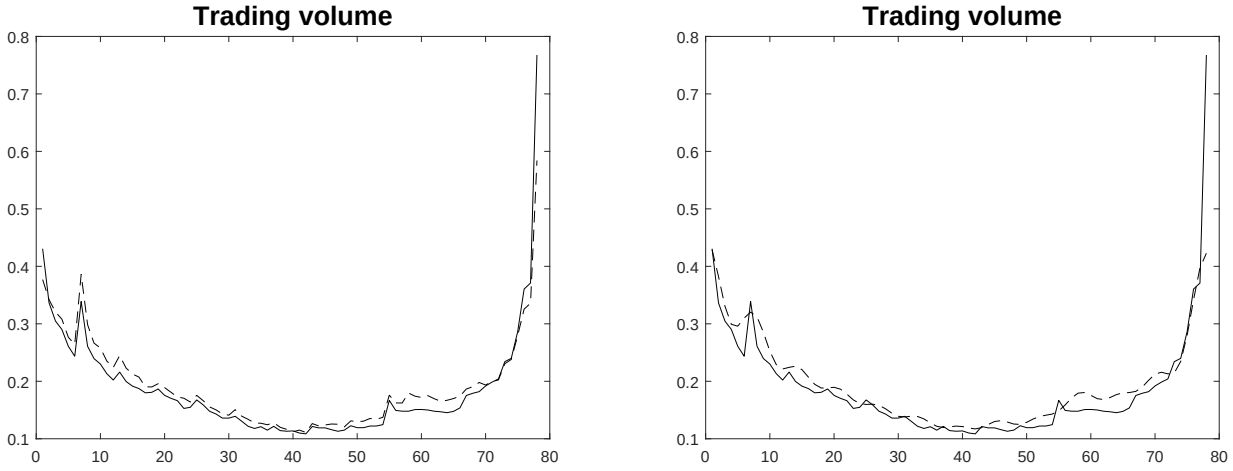
$$\text{MSE}^{VOL} = \sum_{t=1}^{T}\sum_{i=1}^{I}(x_{t,i} - \widehat{x}_{t,i|\cdot}\,)^2 \tag{15}$$

$$\text{MSE}^{VWAP} = \sum_{t=1}^{T}\left(\sum_{i=1}^{I}(w_{t,i} - \hat{w}_{t,i|\cdot}\,)\frac{\bar{p}_{t,i}}{VWAP_t}\right)^2 100^2, \tag{16}$$

in which $VWAP_t$ is defined as

$$VWAP_t = \frac{\sum_{j=1}^{J_t} V_t(j)p_t(j)}{\sum_{j=1}^{J_t} V_t(j)},$$

where $p_t(j)$ and $V_t(j)$ are, respectively, the price and volume of the j-th transaction on day t and J_t is the total number of trades on day t, $w_{t,i}$ is the proportion of volumes traded in bin i on day t, namely $w_{t,i} = x_{t,i}/\sum_{i=1}^{I} x_{t,i}$, and $\bar{p}_{t,i}$ is the VWAP of the i-th bin. Weights $\hat{w}_{t,i|t-1}$ are obtained from predictions $\hat{x}_{t,i|t-1}$ using the one-day-ahead forecasting strategy and weights $\hat{w}_{t,i|i-1}$ are calculated from one-bin-ahead predictions $\hat{x}_{t,i|i-1}$. A comparison of forecasts from the models is made using the Diebold-Mariano test of equal predictability (Diebold and Mariano, 1995).



(a) Linear system (b) CMEM

Figure 3: Forecast of an average trading volume (dashed line) with an actual volume (continuous line) obtained from a one-day-ahead forecasting scheme. The sample is from 21 May 2011 to 30 December 2016.

Table 2: Out-of-sample volume and VWAP tracking forecasting results. For SPY the table reports the values of the volume and VWAP tracking error loss functions. (*) highlights the results of the Diebold-Mariano test at the 1% level.

	MSE^{VOL}	MSE^{VWAP}
One-day-ahead		
CMEM	0.2281	4.1396*
Linear system	0.1932*	4.1766
One-bin-ahead		
CMEM	0.1920	0.2303
Linear system	0.1431*	0.1711*

According to MSE^{VOL} (Table 2), the linear system performs significantly better than CMEM. This is true for both the one-day-ahead and one-bin-ahead strategies. Comparing WVAP tracking errors, the performance of CMEM is better than that of the linear system for the one-day-ahead forecast, but the linear system is significantly more accurate in the one-bin-ahead case. The DM test shows that the forecasts from these models are statistically different in all cases.

4 Conclusion

This paper outlines a multiple-equation approach to forecasting trading volume based on a system of linear equations. It is shown that this system can capture the salient features of volume data, namely the diurnal U-shaped pattern, the periodic, and non-periodic dynamics. The out-of-sample forecast results shows that the linear system performs better than a component multiplicative error model (CMEM) when a simple MSE criterion is used. VWAP forecasting results show that CMEM performs better out-of-sample for the one-day-ahead replication strategy, but the linear model is preferable in terms of a one-bin-ahead forecast. In conclusion, the multiple equation approach is a viable alternative to more complex nonlinear approaches and provides a general framework that can be applied to any high-frequency financial data that exhibits diurnal patterns.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Berkowitz, S.A., Logue, D.E., and Noser, E.A. 1988. The total cost of transactions on the NYSE. *The Journal of Finance*, **43**, 97–112.

Bialkowski, J., Darolles, S., and Le Fol, G. 2008. Improving VWAP strategies: a dynamic volume approach. *Journal of Banking & Finance*, **32**, 1709–1722.

Bollerslev, T., Patton, A.J., and Quaedvlieg, R. 2016. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, **192**, 1–18.

Brownlees, C.T., Cipollini, F., and Gallo, G.M. 2011. Intra-daily volume modeling and prediction for algorithmic trading. *Journal of Financial Econometrics*, **9**, 489–518.

Clements, A.E., Hurn, A.S., and Li, Z. 2016. Forecasting day-ahead electricity load using a multiple equation time series approach. *European Journal of Operational Research*, **251**, 522–530.

Corsi, F. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, **7**, 174–196.

Diebold, F.X., and Mariano, R.S. 1995. Comparing predictive accuracy. *Journal of Business and Economics Statistics*, **13**, 253–263.

Engle, R.F., and Gallo, G.M. 2006. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, **131**, 3–27.

Espinoza, M., Joye, C., Belmans, R., and De Moor, B. 2005. Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series. *IEEE Transactions on Power Systems*, **20**, 1622–1630.

Madhavan, A.N. 2002. VWAP strategies. *Trading*, **2002**, 32–39.

Moisan, S., Herrera, R., and Clements, A. 2018. A dynamic multiple equation approach for forecasting PM2. 5 pollution in Santiago, Chile. *International Journal of Forecasting*, **34**, 566–581.

Naimoli, A., and Storti, G. 2019. Heterogeneous component multiplicative error models for forecasting trading volumes. *International Journal of Forecasting*, **35**, 1332–1355.

Peirson, J., and Henley, A. 1994. Electricity load and temperature: Issues in dynamic specification. *Energy Economics*, **16**, 235–243.

Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F., and Brace, C. 1997. Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, **13**, 161–174.

Soares, L. J., and Medeiros, M. C. 2008. Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *International Journal of Forecasting*, **24**, 630–644.

Spliid, H. 1983. A Fast Estimation Method for the Vector Autoregressive Moving Average Model With Exogenous Variables. *Journal of the American Statistical Association*, **78**, 843–849.