Predicting the Daily Sales of Mikkeller Bars Using Facebook Data

Lisbeth la Cour, Anders Milhøj, Ravi Vatrapu, and Niels Buus Lassen

Journal article (Published version)

Predicting the daily sales of Mikkeller bars using Facebook data

Lisbeth la Cour, Dep. of Economics, CBS
Anders Milhøj, Dep. of Economics, KU
Ravi Vatrapu, Dep. of IT Management, CBS
Niels Buus Lassen, Dep. of IT Management, CBS

1. Introduction.

The present study is a continuation of the analysis presented in Buus Lassen et. al. (2017) in that it still focuses on how to model and predict series of interest to the management of a private firm using social media data. In the present study we focus on only one such data source: Facebook (FB). As mentioned in the paper above: "The main advantage of using social media data as predictors lies in the speed with which such data can be extracted and employed in the forecasting process. Once a firm has learned how to collect and pre-process their social media data, the information is available almost in real time and this implies that such data in combination with a good predictive model will provide a very useful tool for the management of the firm."

The advantage of this year's study is that we now have access to daily observations of the sales in a range of Mikkeller bars of which we have chosen to focus on the bar in Viktoriagade. Hence Mikkeller microbrewery is still our case company. Compared to the monthly data of the paper mentioned above, we have an increased number of observations and we also have the possibility to work in more detail on the lags structure of our models. We still have a high focus on the data preparatory work and we also keep in mind that simple benchmark models that use cheap information are very relevant as competing model specifications.

2. Briefly on the existing literature.

The idea of using social media data as predictors for e.g. company sales is not new. When it comes to model building, various experiments have been conducted and a summary of around 40 articles covering the time period 2005 – 2015 can be found in Buus Lassen et al (2017). For the present purpose the most interesting observations from these studies are that 1) almost 50% of the studies use some kind of regression model as their predictive model, 2) the range of social data types studied seem to cover Facebook, Twitter, Google Trends, Instagram, Tumblr, blogs and Youtube.

Theoretically, the argument for considering social data activity as predictors for sales obtains support from e.g. the AIDA model mentioned in Buus Lassen et al (2014). AIDA means *Awareness, Interest, Desire and Action* and refers to stages in a sales

process. If social media data help increase the attention or can be considered a proxy for attention towards a product then it may also affect the final decision about buying. It is the general perception that more attention will increase sales even if the attention is negative.

When it comes to the specification of a set of predictive models we follow the literature and limit ourselves to the class of dynamic regression models. In these models we will have sales as our dependent variable and the FB data as suggested regressors. Facebook data are polished, because people tend to display success and not failures on this social data. This may imply that FB data has a disadvantage as regressors compared to other social data. Still, FB Likes and FB Posts may provide information that links to consumers awareness and in the end their buying of the product and therefore deserves to be considered as predictors in models of company sales.

3. The data and methodology.

In order to build a predictive model for Mikkeller's sales we use data from Mikkellers accounting system combined with Facebook data. In this analysis we have obtained daily sales data from a number of Mikkeller bars in the Copenhagen area: Viktoriagade, Stefansgade and Torvehallerne (the latter is also a Bottle Shop). The data from the bars are quite ideal for our purpose as they will relate directly to consumption of the product and therefore simplifies the way that we think about the lag patterns in the data. The time span of the study has been limited by our access to historical sales data and covers 2 January 2015 – 30 September 2017. In total we have 1003 observations. In order to perform an out-of-sample forecasting exercise we have held back 3 months of sales data as a tests sample while we select and estimate our model based on the remaining around 900 observations.

Prior to analysis we index the sales data such that the mean is restricted to 1234 and the standard deviation to 12. Such transformations do not affect the significance our results later in the modeling process. The Facebook data comes from the overall HQ Mikkeller FB page

https://www.facebook.com/mikkeller/
https://www.facebook.com/events/)

and from the FB pages of the chosen bars

https://www.facebook.com/mikkellerbarvik/,
https://www.facebook.com/MikkellerandFriendsBottleShop/,
https://www.facebook.com/mikkellerandfriends/).

Using the Sodato software developed by Ravi Vatrapu and his group, see Hussain & Vatrapu (2014), we collect information from the selected FB pages and we create variables for e.g. total likes of the posts on a specific date. As the data is very rich, for a shorter sample period we are also able to construct explanatory factors based on selected FB reactions which are constructed to match major human emotions and in this way seems ideal when sales are in focus.

3.1 Pre-processing methodology

Our first considerations when it comes to data preparatory work concerns whether to use simple transformations of the series or just the raw series themselves. As the values of sales are quite low on certain dates it does seem like a disadvantage rather than an advantage to use a log-transformation. Also no clear pattern of an increase in volatility over time is revealed from e.g. Figure 1A and we decided to model the un-transformed series directly.

With respect to the sales data we are checking the stationarity properties of the time series by means of several graphs: sales against time and ACF. We also perform ADF tests of the null of non-stationarity. Stationarity is preferable for a regression model although stationarity may be of minor importance when the purpose of the model is forecasting.

The social data may consist of different components that we would expect to have different predictive value. Prior to including our social data time series as explanatory factors in our regression models we have the possibility to split them into a trend component, a seasonal component and an irregular component using classical times series techniques for unobserved components models (ucm). We also estimate models that use the social data in their 'raw' form without the ucm pre-processing for comparison reasons.

3.2 Unobserved Component Models

We use the same modeling strategy as in Buus Lassen et al (2017) and therefore start out by employing an unobserved component (UCM) model. An UCM decomposes the observed series y_t into a sum of many components, as for instance

$$y_t = \mu_t + \varepsilon_t$$

$$\mu_t = \mu_{t-1} + \eta_t$$

Here the series μ_t is understood as the level of the series; but this level is unobserved. Only the series y_t which is affected by some noise or irregularities is observed. This noise series, ε_t, could in technical applications be measuring errors.

This basic formulation could be extended by trends and seasonality, and various forms for introducing autocorrelation in the model formulation also exist. A trend component

is insignifikant for the sales series. A seasonal component for the day of the week effect is defined in a way so it does not affect the level component:

$$S_t = -(S_{t-1} + ... + S_{t-6}) + \zeta_t$$

In total these ideas lead to the model:

$$y_t = \mu_t + S_t + \varepsilon_t + \varphi_1\varepsilon_{t-1} + \varphi_7\varepsilon_{t-7}$$

where we also include first lag and lag 7 autoregressive terms are included for this series of daily observations with a significant weekly pattern.

All remainder terms, ε_t, η_t, and ζ_t, are assumed to be mutually independent white noise series. Their variances could be estimated; the larger this component variance the more volatile the component. But it is also possible to fix this variance to the value zero which gives a constant component, e.g. a model with fixed seasonal dummies is found if $\text{var}(\zeta_t) = 0$.

The parameters of these models, the variances and the autoregressive parameters, could be estimated by the Kalman filter together with all and the component values. This gives an algorithm for successive calculation of the unobserved components at time t conditioned on previous observations y_{t-i} $i = 0, .. , t-1$. The Kalman filter is useful if prediction is the purpose of the analysis as the algorithm does not include future observations y_{t+i}. A further smoothing estimation, where all available information is used when estimating the unobserved components at any time t, also exist. In this paper this method will be used.

3.3 The regression models

In this study we use dynamic regression models. With daily data we have a rich seasonsal structure and even though we only have a sample period covering less than three years we have enough observations to model the seasonality either by ucm (mentioned earlier) or by inclusion of deterministic dummies in the regression equations. Using lags of both the dependent variable and the independent variables is also possible and we will do both.

The primary model equations we use are of the type:

$$(1) \qquad y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_h y_{t-h} + \beta_1 \Sigma x1_{t-i} + \ldots + \beta_k \Sigma xk_{t-i} + \varepsilon_t \qquad t = 1, .., T$$

where y is sales, the x's are FB measures and the sub scripts, t - i, indicate that only lagged values of sales and FB data are used as predictors. This makes the model suitable for at least 1 step ahead predictions out-of-sample. In practice we use both short lags and lags up to 8 to cover a same-day-of-the-week effect and also an interaction of short run and day-of-week effects. The error term, ε_t, is assumed to fulfill the standard assumptions for OLS estimation.

It is difficult to judge the predictive performance of a specific forecasting model unless we have some benchmark to compare to. For sales of individual companies there is no general guideline in the literature on how to choose such a model, so we will argue for our choice in the following way: we want a benchmark model that is simple, that seem to capture some of the apparent time series properties in our data and that do not contain FB explanatory factors. We choose two benchmark models. The first includes only deterministic terms and a trend:

(2) $y_t = \beta_0 + \beta_1 D0101_t + \beta_2 D2412_t + \beta_3 D2512_t + \beta_4 D2612_t +$ day-of-week dummies + monthly dummies + CBC dummies + trend + ε_t $\quad t = 1, .., T$

The second includes in addition to all the deterministic dummy and the trend and up to 8 lagged values of sales:

(3) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma y_{t-8} + \beta_1 D0101_t + \beta_2 D2412_t + \beta_3 D2512_t +$ $\beta_4 D2612_t$ + day-of-week dummies + monthly dummies + CBC dummies + trend + ε_t $\quad t = 1,.., T$

Finally, as our model is a forecasting model, we need to split the sample into an estimation part and a part used to evaluate the out-of-sample forecasting properties of the model. For further discussion, see e.g. Hyndman & Athanasopoulos (2014). We retain the last 3 month of the sample for the test part, i.e. July – September 2017 (92 observations) and we provide 1-step-ahead prediction for this period. Hence we estimate the models using data from 1 January 2015 until 30 June 2017 (around 904 observations). When we use the FB reaction we stick to the same evaluation sample but we have a shorter estimation sample as the FB reactions were introduced in the beginning of 2016. Evaluations will be based on graphs comparing actual sales to predicted sales as well as by numerical measures like RMSE and MAE.

4. Descriptive statistics.

We start by showing some graphs and descriptive statistics for the sales data. In Figure 1A we show the development over time in the standardized sales at Viktoriagade Bar over the sample period[1]. The immediate impression is a series that do not show a trending behavior. There are three cases of very large sales in certain spring weekends coinciding with the Copenhagen Beer Celebration . Also some seasonal variation can be seen. To illustrate the over-the-week pattern in the series we have constructed the special graph shown in Figure 1B. We selected (randomly) 6 consecutive weeks during the summer of 2015 (15 June –26 July). Each curve in Figure 1B shows the

[1] On the 1st January each year the bar is closed and numbers for sales are missing. Instead of filling in zeros at this stage we simply do not show these dates in the graphs. When modeling we add dummies to capture these dates without any sales.

sales for a week during this period. The figure displays a pattern of larger sales on Fridays and Saturdays and also to some extent that the level of the sales may depend on the week (maybe the weather – maybe vacation weeks). Taking this intra-week pattern into account will also be important for our modeling.

Figure 1A: Sales, Viktoriagade

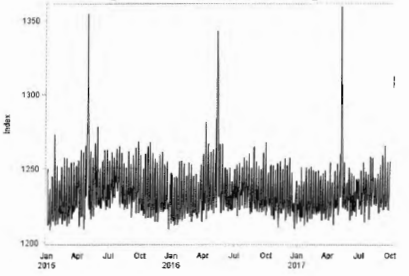
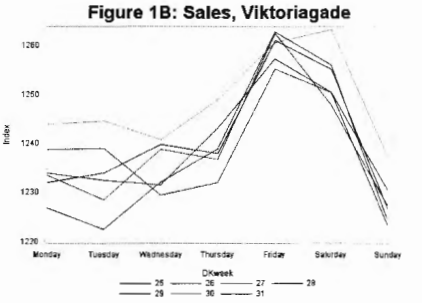


Figure 1B: Sales, Viktoriagade



When taking a closer look at the time series properties of the series it seems that a decision of treating this series as stationary would be a good starting point. The ACF graph of the sales corrected for the missing sales of 1 January seems to support this conclusion as the 1st order autocorrelation coefficient is 0.55[2]. See also Figure 2A.

Figure 2A: ACF for Sales 1janD

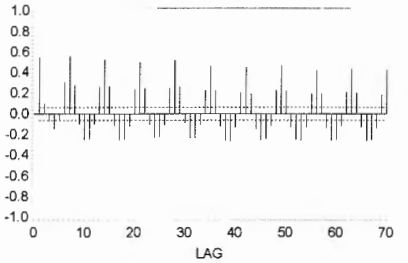
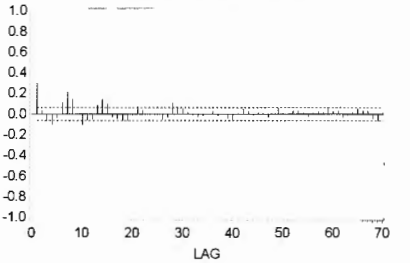


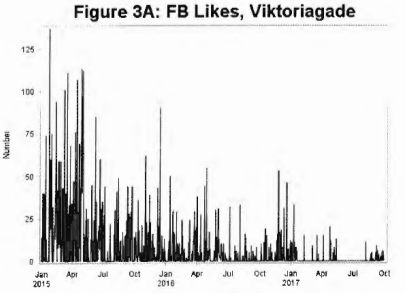
Figure 2B: ACF for Sales more D's

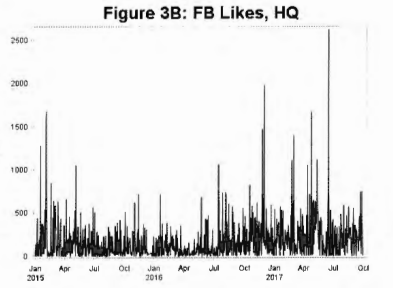


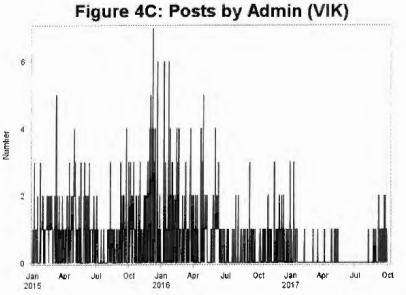
In this graph we also see clear indications of an over-the-week pattern. In Figure 2B we have in addition to correcting for the 1 January also regressed sales on dummies for each day-of-the-week, each month and for dates around Xmas and the Copenhagen Beer Celebration event (the three large spikes in Figure 1A). This implies that the memory of the seasonal pattern becomes less pronounced. Inspired by the PACF of the extended model (available from the authors upon request) the model can be extended by 8 lagged values of sales and after such an extension almost no autocorrelation is left.

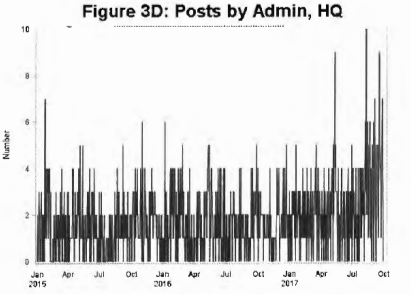
[2] Also an ADF test of non-stationarity of the series supports a conclusion of stationarity. With an intercept, but without a trend in the equation of this test we reject at the 1% level the null of a unit root with p-values smaller than 0.0001 and 0.0001 for zero and 7 lagged differences in the equation, respectively.

To get a first impression of the some of the data from Facebook, in Figures 3A – 3D shows the number of likes and the number of posts by the administrator for the HQ page and for the Viktoriagade page. While none of the series seem to follow the pattern of the sales series very closely they seem to correlate pairwise (Viktoriagade – Viktoriagade and HQ – HQ). Also the number of Likes for HQ are - not surprisingly - in general larger than for Viktoriagade. Notice that the activity for Viktoriagade show a decline in 2017 compared to the other years (Mikkeller has no specific explanation to that).









In Figure 4 we look for correlations between sales and the selected FB variables. Correlations amongst the FB variables are also shown. It is not easy to get a clear idea about the relationships. There may be indications of weak positive correlations in most of the cases but as we may want to use the FB variables from a range of previous days as regressors the scatterplot matrix is not the best graph for that purpose. In Figure 5 we show a selected cross-correlation graphs to get a better idea of a potential lag-pattern.

Table 1 shows simple descriptive statistics for the variables we have been investigating so far. The numbers of the mean and standard devistion for sales reflect our standardization. The three missing values are 1 January each of the three years. Not surprisingly we see both more posts and also more reactions to the HQ FB activity.

Table 1: Descriptive summary statistics.

Variable	N	Mean	Std Dev	Minimum	Maximum	N Miss
Sales	1001	1234.02	17.01	1209.1	1359.42	3
Posts by Admin (VIK)	1004	0.57	0.95	0	7	0
Posts by Admin (HQ)	1004	1.74	1.48	0	10	0
Likes by Viktoriagade	1004	6.71	15.70	0	137	0
Likes by HQ	1004	168.55	230.94	0	2626	0

Figure 4: Scatterplot Matrix for Viktoriagade Sales

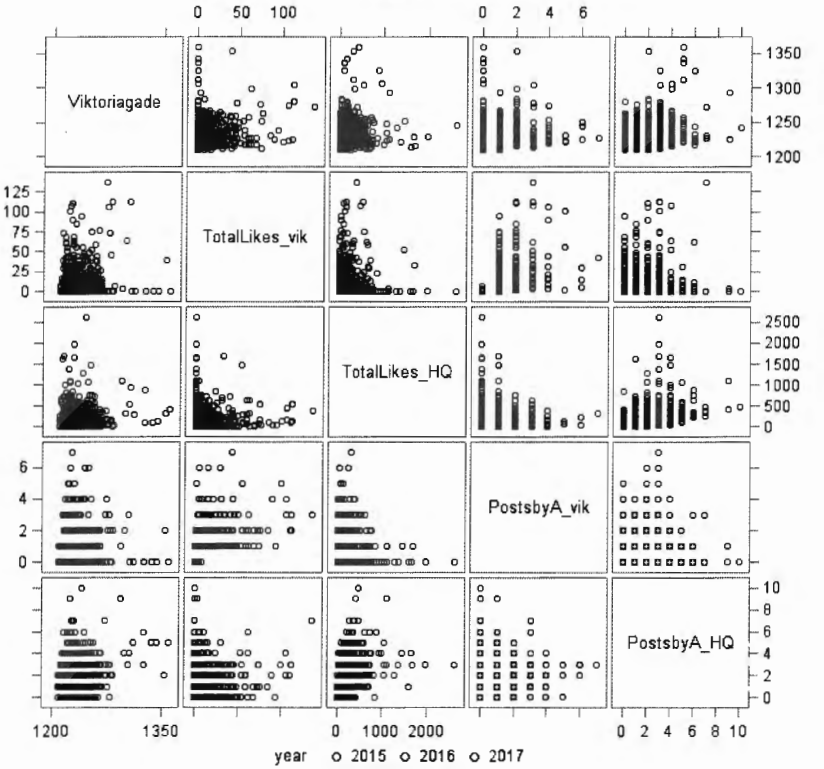
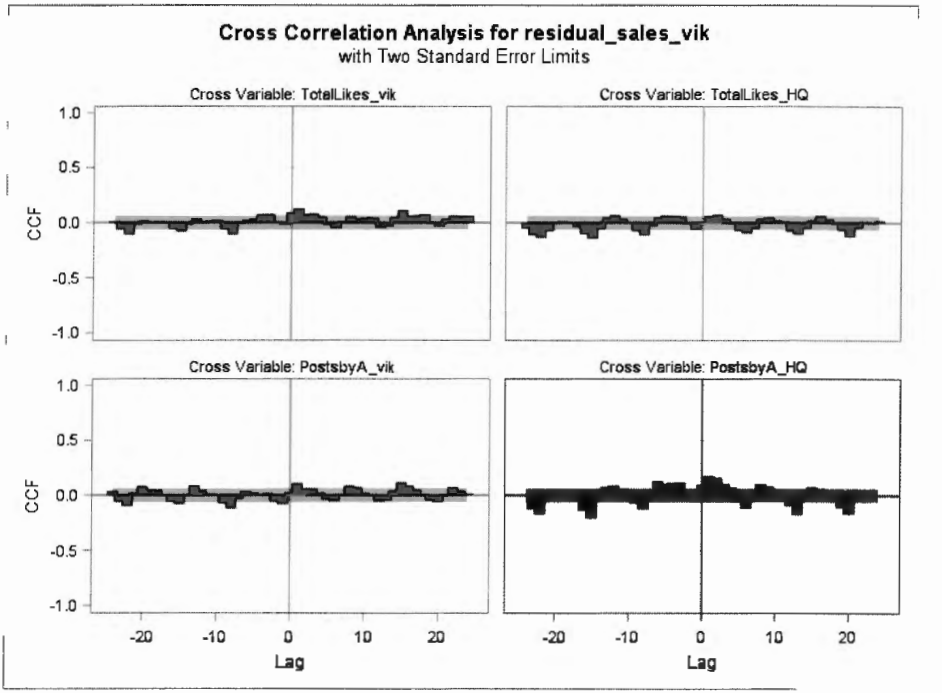


Figure 5 presents the cross correlations between the indexed sales variable and each of the four social activity variables. The cross correlations are constructed in such a way that a positive number, s, on the horizontal axis implies a correlation between sales at time t and the social variable s periods prior to time t. Even though we did not pre-whiten the series before the cross correlations were calculated we get some initial impression that lagged values of both likes and posts may contain explanatory power for the sales.

Figure 5:



5. Unobserved components models for the sales series

For the sales series the resulting model is:

$$y_t = \mu_{t-1} + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$$

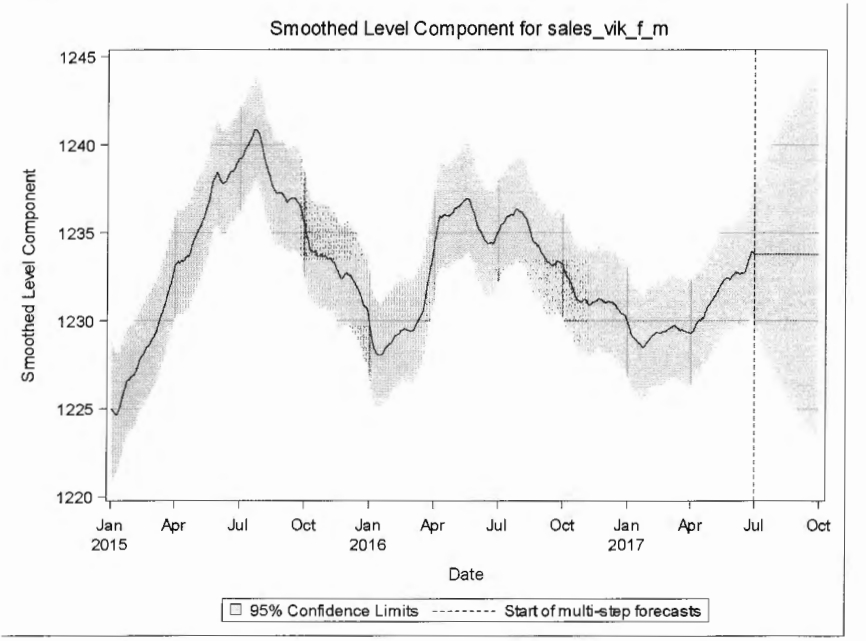
The variance in the seasonal component is fixed to zero, meaning that the dummy variables are constant. It turns out to be inconvenient to model annual variation in the sales by cyclic components. Instead the level component is modeled with a positive

component variance, var(η_t) > 0. Figure 6 shows the resulting estimated level component.

The final version of this model is estimated without dates for Copenhagen Beer Celebration, some days around Christmas and January 1'st where sales for well-known reasons are extraordinary. In the set up for UCM models this is done by simply setting the observations to "missing" instead of using dummy variables. But even some more dates give clear outliers in the fitted models. For this reason, it is chosen to also set five more observations as missing because they give clear outliers. It was checked that the sales these dates could not be explained by our Facebook data so the outliers must be due to something else - perhaps some extraordinary event in the bar.

The precise choice of the number of outliers to leave out is of course subjective, but it has to be stressed that the validity of an out-of-sample forecasting exercise is independent of the number of observations left out from the estimation period.

Figure 6



The level varies around the average value 1234 - remember that data is standardized to this mean. The estimated component variance is var(η_t) = 0.28. This level variance gives clearly visible changes in the level, but only in an interval ± 10 - remember that the series is standardized to variance standard deviation 12.

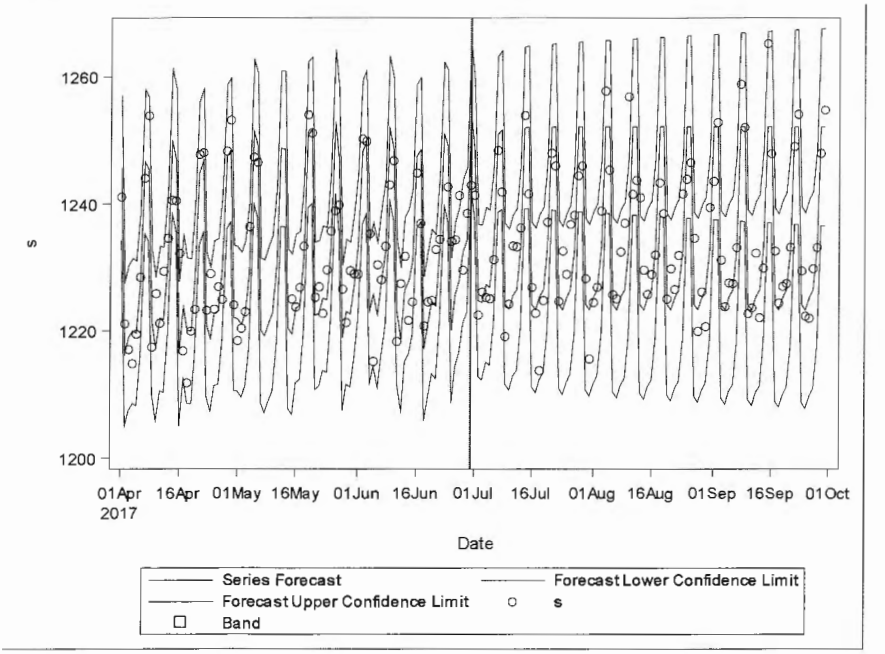
The level of the series is highest in the summer period but the annual variation is far from regular, so this more flexible model for the annual variation could be superior to monthly dummy variables or cyclic components.

The final model also includes dummy variables for the weekly effect. It was tried model the seasonal component with a time varying weekly pattern but the hypothesis that the component variance was zero, var(ζ_t) = 0, was accepted - however borderline p = 6.9%. The main feature for a potential time varying weekly pattern is that the Friday effect of a is reduced from 24 to 19 in the scale used.

5.1.2 Out-of-sample predictive power?

This model without any exogenous variables could by applied for forecasting. Figure 7 shows the results. For the last quarter of the estimation period - that is April 1'st 2017 to Jun 30'th 2017 the plot gives one step ahead predictions with forecast limits as opposed to the actual observations. For the last quarter - July 1'st 2017 to September 30'th 2017 the predictions are made using only data and estimation results before the last date of the estimation period - that is June 30'th as indicated by the vertical reference line. For the 92 observations in this ex-ante forecasting period, July 1'st 2017 to September 30'th 2017, we find RMSE = 6.36 and MAE=4.83.

Figure 7



6. Results of predictive modeling at the daily frequency

We now consider various specifications for models that contain FB data and/or their lags as explanatory factors as suggested by the main equation (1). As our main purpose is to determine a model that can produce out-of- sample 1 step ahead forecasts, we do not use contemporaneous regressors in the models.

6.1.1 Estimation results, regression models

Estimation results for a selected range of regression models are shown in Table 2. To save space we have just commented on the results of most of the dummies without actually showing the coefficients. For all Xmas dummies the coefficients are negative. For the day-of-week dummies the coefficients are always positive indicating that Mondays in general have the lowest sales (the base category) while sales are highest on Fridays and Saturdays (not surprisingly). For the monthly dummies the base category is January and the sales in all other months are significantly higher than for January and most so for May until September. For Copenhagen Beer Celebration the sales are in general higher and very much so when getting closer to the weekend.

The basic massage from Table 2 is that it is very hard in-sample to beat a model with just deterministic terms as explanatory factors as our Benchmark 1. Only in one version do we find significance of any of the FB variables and those results are the ones shown in the last column of Table 2. Here the likes of HQ at lag 1 and at lag 7 are significant although with coefficients of a sign opposite to the expected one. To move from the full model to the model with just 2 Likes-variables included we did a range of F tests for exclusion of likes and posts variables.

6.1.2 Out-of-sample predictive power?

We predict the standardized sales for the time period July 2017 to September 2017. First we show graphs, Figure 8, that compares such predictions to the actual values. We show graphs for the benchmark model 1 and for the model in the last column of Table 2.

From these graphs it is evident that most of the movements in sales are captured by the benchmark model. The confidence bands for the prediction are quite wide, however, indicating a fairly high uncertainty for the forecasts. Most of the actual values are inside the bounds except for 2 incidents in mid-July and mid-September. The picture shown for the model with lag 1 and lag 7 of Likes of HQ is very similar.

Table 2: Regression results for Log Sales - no ucm.

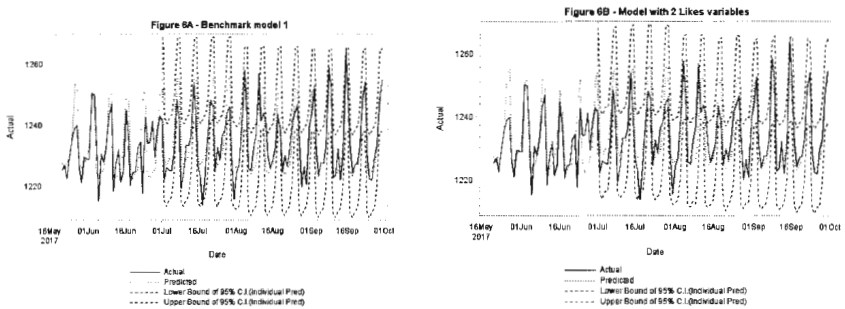
Variables	Benchmark1 det. terms	Benchmark2 AR(8) and det.	Only lagged Sales AR(8)	Full Model Equation (1)	Model with all det. and sign. likes
Intercept	Significant	Significant	Significant	Significant	Significant
Xmas D's	Significant	Significant	-	Significant	Significant
Week day D	Significant	Significant	-	Significant	Significant
Monthly D's	Significant	Significant	-	Significant	Significant
CBC D's	Significant except for Monday and Tuesday in 16 and 17	Significant except for Monday and Tuesday in 16 and 17	-	Significant except for Monday and Tuesday in 16 and 17	Significant except for Mondayand Tuesday in 16 and 17
Trend	-0.004*** (0.001)	-0.004*** (0.001)	-	-0.005*** (0.001)	-0.003*** (0.001)
Sales, lag1	-	-0.001 (0.004)	0.041*** (0.009)	-0.000 (0.004)	-
Sales, lag6	-	0.006 (0.004)	0.027*** (0.009)	0.007* (0.004)	-
Sales, lag7	-	0.005 (0.004)	0.047*** (0.009)	0.007* (0.004)	-
Sales, lag8	-	0.004 (0.003)	0.014* (0.008)	0.005 (0.003)	-
Sales lags 2-5	-	Insignificant	Insignificant	Insignificant	-
Lags 1-8 VIK Likes	-	-	-	Insignificant	-
Lags 1-8 VIK Posts	-	-	-	Insignificant	-
Lag 1, HQ Likes	-	-	-	-0.003** (0.001)	-0.003*** (0.001)
Lag 7, HQ Likes	-	-	-	-0.003** (0.001)	-0.003*** (0.001)
Lags 2-6, 8 HQ Likes	-	-	-	Insignificant	-
Lags 1-8 HQ Posts	-	-	-	Insignificant	-
Adj. R square	0.987	0.922	0.987	0.988	0.988
#observations	904	904	904	904	904

Note: the estimation sample has been restricted such that it is the same for all specifications even though models with fewer lags could have used more observations.

Note2: Standard errors in parentheses. Significance at 10%: *, 5%: **, 1%: ***.

Note3: Dummy for 1 January included and significant in all models.

Figure 8



In Table 3 we show some numerical measures for the forecasting performance of the models from Table 2. We have chosen just to focus on a few measures and some of the more commonly used ones: MAE (mean absolute error) and RMSE (root mean squared error)[3].

Table 3: Summary measures on predictive power.

Summary measure	Benchmark1 (det. terms)	Benchmark2 AR(8) and det.	Only lagged Sales AR(8)	Full Model (1)	Model with all det. and sign. likes
MAE	5.06	5.08	8.57	5.22	5.13
RMSE	6.73	6.74	10.34	6.85	6.79

Note: In all cases the numbers have been calculated based on the 3 months of July, August and September 2017.

The numbers in table 3 also indicate that benchmark model 1 performs the best both when evaluated based on MAE and on RMSE. However only the model with just the lagged sales variables (the middle column) shows somewhat higher statistics. The other numbers are actually very close. We have not performed a formal test of equality.

Notice that our forecasting period does not contain the week of CBC. As it stands our models are not well suited to forecast for a time period that contains this week as we would then have to come up with predictions for the excess sales of that week (in e.g. 2018). In future work we could have restricted the coefficients of the CBC dummies to always be the same and in this way have handled that problem.

[3] For formulas on how to calculate these measures, please consult e.g. Hyndman & Athanasopoulos (2014)

6.2 Facebook data used in the Unobserved Components Model

6.2.1 Estimation results using regression methods

This approach this year is different from the approached by Buus Lassen et al (2017) were the input variable were applied to the irregular series as extracted by the UCM.

First only lagged values of the FB data was used to predict sales. All four series of FB observations, Posts by Admin (Viktoriagade) Posts by Admin (HQ), Likes by Viktoriagade and Likes by HQ, were used with lags from 1 to 8. This total of 32 exogenous variables were used as ordinary input variable to the UCM found i Section 5. All regression coefficients and all parameters and component values in the full model were estimated simultaneously.

As in the OLS models in section 6.1 most of the input variables are insignificant. Table 4 gives two significant regression coefficients along with the parameters of the UCM. The significant coefficients are lag 1 of the number of total likes for HQ; the coefficient however has a negative sign, which is in contrary t our intuition, The second coefficient is for lag 1 of the number of posts from the specific bar in question, Viktoriagade. The coefficient tells that for each post from the Viktoriagade bar the sales next day at the Viktoriagade bar increases by 0.47 in our scaled sales. This is a result that has a potential for active marketing. The number 0.47 is however small when compared to the average daily sales, which was set to the number 1234.

Table 4

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	31.19381	1.57766	19.77	<.0001
Irregular	AR_1	0.28156	0.03712	7.59	<.0001
Irregular	SAR_1	0.13395	0.03679	3.64	0.0003
Level	Error Variance	0.26927	0.10766	2.50	0.0124
ltotallikes_hq	Coefficient	-0.00189	0.0008336	-2.27	0.0235
lpostsbya_vik	Coefficient	0.47150	0.20144	2.34	0.0192

This model could be applied i the out-of-sample forecasting exercise. This gives MAE = 4.84 and RMSE = 6.34. These values are very close to the values obtained without using the FB data.

When also unlagged observations of the four input series are used one more input variable shows significance; see Table 5. The unlagged total likes for Viktoriagade bar

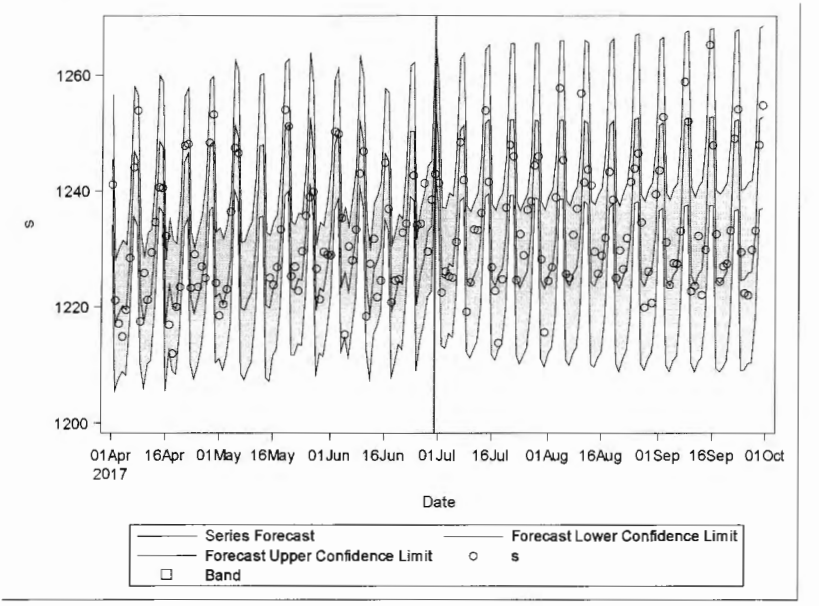
has the coefficient 0.041 - meaning that each like corresponds to an increasing sale of 0.041 beers in our scale. However, this effect is probably a reverse causal effect as many likes one evening probably leads to more instantaneous likes.

Table 5

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	30.85263	1.56438	19.72	<.0001
Irregular	AR_1	0.29117	0.03736	7.79	<.0001
Irregular	SAR_1	0.13139	0.03693	3.56	0.0004
Level	Error Variance	0.28051	0.11220	2.50	0.0124
TotalLikes_vik	Coefficient	0.04135	0.01337	3.09	0.0020
ltotallikes_hq	Coefficient	-0.00177	0.0008287	-2.13	0.0328
lpostsbya_vik	Coefficient	0.60043	0.20406	2.94	0.0033

Figure 9



When this model is used in the out-of-sample exercise we find MAE= 4.83 and MSE = 6.32. Again these values are very close to the values obtained without using the FB data and the model only using unlagged input variables.

7. Summary and conclusion

In this paper we have pursued our idea of applying a preparatory ucm model to both regressors and regressand to determine a forecasting model for the monthly sales of the Danish microbrewery Mikkeller. Also we tried a more traditional strategy with lagged sales to model the autocorrelation of the in the sales series and a suite of dummy variables for deterministic outside factors; take the effect of Xmas as an example in order to build a predictive model.

Our modeling attempts were mainly unsuccessful as neither of the two approaches lead to any significant regression model when Facebook activity was included as input variables.

8. References

Buus Lassen, N., la Cour, L., Milhøj, A., Vatrupu, R. (2017), 'Social media data as predictors of Mikkeller sales?' in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 71-86

Buus Lassen, N., la Cour, L., Vatrapu, R. (2017), 'Predictive Analytics with Social Media data' in Sloan & Quan-Haase ed. *The SAGE Handbook of Social Media Research Methods*, Chapter 20, pp 328-341

Buus Lassen, N., Madsen, R. and Vatrapu, R. (2014). 'Predicting iPhone Sales from iPhone Tweets', Conference Paper, *2014 IEEE International Enterprise Distributed Object Computing Conference.*

Buus Lassen, N., Vatrapu, R., la Cour, L., Madsen, R. and Hussain, A.(2016), 'Towards a Theory of Social Data: Predictive Analytics in the Era of Big Social Data', in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 241-256

Doornik & Hendry (2014). 'Statistical Model Selection with 'Big Data', *Department of Economics Discussion Paper Series*, University of Oxford, #735.

Hussain A., Vatrapu R. (2014) Social Data Analytics Tool (SODATO). In: Tremblay M.C., VanderMeer D., Rothenberger M., Gupta A., Yoon V. (eds) Advancing the Impact of Design Science: Moving from Theory to Practice. DESRIST 2014. Lecture Notes in Computer Science, vol 8463. Springer, Cham

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*: OTexts: https://www.otexts.org/fpp/