# Forecasting High-Frequency Electricity Demand with a Diffusion Index Model*

Gabriel Pons Rotger

*Institute of Local Government Studies, Denmark*

gpr@akf.dk

Philip Hans Franses

*Econometric Institute, Rotterdam*

franses@few.eur.nl

Econometric Institute Report 2006-38

**Abstract**

We propose a diffusion index model (Stock and Watson, 2002) to forecast electricity demand for one hour to one week ahead. The model is particularly useful as it captures complicated seasonal patterns in the data. The forecast performance of the proposed method is illustrated with a simulated real-time experiment for data from the Pennsylvania-New Jersey-Maryland Interchange.

Key words and phrases: Diffusion Index Forecast; Seasonality; Electricity Load.

JEL CODES: C12, C22, C32.

*Gabriel Pons Rotger appreciates support from the Danish Social Science Research Council. Corresponding author: Philip Hans Franses at Econometric Institute, Erasmus University Rotterdam, P.O.Box 1738, NL-3000 DR Rotterdam

# 1 Introduction

This paper is concerned with the problem of forecasting electricity demand in the short run, where the focus is on hours within weeks. This type of forecast provides the input of scheduling algorithms for the generation and transmission of power, and therefore is a fundamental tool to improving load dispatching, unit commitment and system stability. In addition, with the advent of deregulated electricity markets around the world, load forecasting has become even more relevant since unexpected changes in electricity demand is a major input to forward electricity price premia (see Bessembinder and Lemmon, 2002).

Electricity load, in contrast to electricity prices, is a relatively easy-to-forecast time series. However, the associated cost of load forecast errors is so high, that even a slight reduction of forecast error has very important consequences for the many agents involved in the electricity market, and, subsequently of course, for the consumers of electricity.

The most common methods to forecast short-run demand are seasonal autoregressive and moving average models, exponential smoothing methods (see Taylor et al, 2006), artificial neural networks, moving averages, multi-equation models, or periodic splines (see Harvey and Koopman, 1993). However, a common feature of these forecasting methods is that they do not properly account for an important stylized fact of the data, and that is the presence of periodic autocorrelation, which formally means that

$$\begin{aligned} E(y_{it}y_{i-kt}) &\neq E(y_{i+mt}y_{i+m-kt}),\ i=1,...,N, \\ E(y_{it}y_{i-kt}) &= E(y_{i+Nt}y_{i+N-kt}). \end{aligned} \tag{1}$$

where $N$ is the number of seasons. Furthermore, due to the non-storability of power, its demand is strongly determined by temperature, which is also a periodically correlated time series, see Lund et al. (1995). Finally, the habits of electricity end-users could impart periodic autocorrelation as well (Hansen and Sargent, 1993). Hence, to model electricity demand, one needs to capture the periodic features in the data.

A general class of time series models that generates (1) concerns the periodic autoregressive and moving average (PARMA) family (see Tiao and Grupe, 1980), given by

$$(y_t - \mu_i) - \Sigma_{k=1}^{p_i} \phi_{k,i}(y_{t-k} - \mu_{i-k}) = \Sigma_{k=0}^{q_i} \theta_{k,i} \eta_{t-k},\ i = 1, ..., N \tag{2}$$

where $\eta_t$ is a zero-mean process with unit variance and where the parameters $\mu_i$, $\phi_{k,i}$, $\theta_{k,i}$ and the model orders $p_i$, $q_i$ are allowed to vary with the season $i$. In model (2) the different cyclical components are not mutually orthogonal, that is, the intra-day components are correlated with the intra-week seasonality, and consequently, seasonal adjustment methods cannot completely isolate the components.

When the number of seasons is small, the $2N + \Sigma_{i=1}^{N}(p_i + q_i)$ parameters of (2) can be estimated with methods adapted to the periodic model (see Pagano, 1978; Anderson and Vecchia, 1993). However, even for moderate $N$, PARMA model parameters can only be estimated for small orders $p_i$, $q_i$ and under smooth periodicity of $\phi_{k,i}$ and $\theta_{k,i}$. Misspecification and the absence of reasonable restrictions on (2) is likely to be the main reason for the sometimes observed poor forecasting performance of periodic models (see Franses and Paap, 2004).

The important contribution of the seasonal component to the dynamics of electricity demand suggests that a proper description of (1) may improve the forecast performance with respect of the existing methods. In this paper we attempt to do so. Our method exploits principal components and regression to solve the dimensionality problem implied by $N \to \infty$ and (1). We propose to extract from the data latent factors that drive most of the dynamics, and we use them in a diffusion index (DI) forecast (see Stock and Watson, 2002). The DI model has been relatively successful to forecast macroeconomic time series, and in this paper we demonstrate that such a method also produces quite good short-run forecasts of high-frequency electricity demand.

The outline of our paper is as follows. Section 2 describes the DI model for our data, where basically the cross-sections correspond to different seasons (here: hours within a week) of the same underlying time series. Section 3 discusses the analysis of unit roots in the framework of factor analysis and shows the implications for the underlying periodic series. Section 4 describes the estimation and specification of the DI model. Section 5 presents the point and interval DI forecasts. Section 6 illustrates the performance of the DI model with a simulated real-time experiment with hourly data from the Pennsylvania-New Jersey-Maryland Interchange (PJM) Mid-Atlantic Region (January 7, 2002-June 18, 2006). Section 7 concludes.

# 2 The model for hours within a week

Let $y_{it}$ be a stochastic variable measured at hour $i$ within week $t$ and denote $\mathbf{y}_t = (y_{1t}, ..., y_{Nt})'$ as the vector of all periods $i$ within that interval, see Franses (1994). In our application where we consider hourly electricity demand, $\mathbf{y}_t$ is a vector composed by the load at the 168 different hours of week $t$.[1]

Suppose that for $h = 1, ..., N$, ($\mathbf{y}_t$,$y_{ht+1}$) admits the dynamic factor model representation (see Forni et al., 2000), that is,

$$\begin{aligned} y_{ht+1} &= \upsilon_h'(L)\mathbf{f}_t + \beta_h'\mathbf{W}_{ht} + \varepsilon_{ht+1},\ t = 1, ...T-1 \quad (3) \\ y_{it} &= \lambda_i'(L)\mathbf{f}_t + \xi_{it},\ i = 1, ..., N;\ \ t = 1, ..., T \quad (4) \end{aligned}$$

where $\upsilon_h(L) = \Sigma_{j=0}^{s+1}\upsilon_{hj}L^j$ and $\lambda_i(L) = \Sigma_{j=0}^{s}\lambda_{ij}L^j$ are lag polynomials of finite order, $\mathbf{f}_t = (f_{1,t}, ..., f_{q,t})'$ is a $q$-dimensional unobserved set of dynamic factors with $q << N$, $\xi_{it} \equiv y_{it} - \lambda_i'(L)\mathbf{f}_t$ is the idiosyncratic error, $\mathbf{W}_{ht}$ is a vector of observable predictors for the idiosyncratic component of $y_{ht+1}$, like lags of periods similar to $h$, and where $\varepsilon_{ht+1}$ is the error of the forecasting equation.

The dimensions of the panel $N$ and $T$ are assumed to be large. In practice the electricity load variable is generated continuously and available at a high frequency, like at 5 minutes, 30 minutes, hours and so on. The large $N$ framework suits in our case where the load is measured as frequent as the hour.

The component $\lambda_i'(L)\mathbf{f}_t$ captures the common dynamics of the $y_{it}$ variables. The factors $\mathbf{f}_t$ may be persistent and or have short memory. The

[1] For example $y_{1t}$ denotes the load of Monday at 1 AM of week $t$, and $y_{168t}$ the load of Sunday at 12 PM of week $t$.

idiosyncratic component $\xi_{it}$ captures the specific dynamics of subsets of periods like load peak periods or weekends.

The dynamic relation between $\mathbf{f}_t$ and $y_{it}$ reflects the presence of co-movements across periods pertaining not only to the same week but also to consecutive weeks. Our model (3)-(4) can be written in static form like

$$y_{ht+1} = \Upsilon_h' \mathbf{F}_t + \beta_h' \mathbf{W}_{ht} + \varepsilon_{ht+1},\ t = 1, ..., T-1 \tag{5}$$

$$y_{it} = \mathbf{\Lambda}_i' \mathbf{F}_t + \xi_{it},\ i = 1, ..., N;\ t = 1, ..., T \tag{6}$$

where $\Upsilon_h = (\upsilon_{h0}', ..., \upsilon_{hs+1}')'$, $\mathbf{\Lambda}_i = (\lambda_{i0}', ..., \lambda_{is}')'$, and where $\mathbf{F}_t = (F_{1,t}, ..., F_{r,t}) \equiv (\mathbf{f}_t', \mathbf{f}_{t-1}', ..., \mathbf{f}_{t-s}')'$ with $r = q(s+1)$.

For this model (5)-(6), we adopt the following assumptions:

*Common Component*

1.1 $E(\mathbf{F}_t \mathbf{F}_t') = \Sigma_{FF}$ where $\Sigma_{FF} = diag(\sigma_{11}, ..., \sigma_{rr})$ with $\sigma_{ii} > \sigma_{jj} > 0$ for $i < j$ and $T^{-1} \Sigma_{t=1}^{T} \mathbf{F}_t \mathbf{F}_t' \xrightarrow{p} \Sigma_{FF}$.

1.2 The loadings $\Lambda_{ji}$ satisfy $N^{-1} \mathbf{\Lambda}' \mathbf{\Lambda} \xrightarrow{p} \Sigma_{\Lambda}$ with $\Sigma_{\mathbf{\Lambda}}$ an $r \times r$ nonrandom matrix and $|\Lambda_{ji}| \leq \overline{\Lambda} < \infty$ for all $j$, $i$.

The dynamic factors have constant unconditional second moments. The loadings vary across $i$, that is, $\mathbf{\Lambda}_i \neq \mathbf{\Lambda}$, and this permits that the common shocks have different impact on different periods within the cross-section, here hours. This is very likely in our application where the proportion of residential and non-residential electricity end-use presents a very important variation within the day and week. In fact, this feature generates stronger correlations among periods with similar electricity end-use composition.

*Idiosyncratic Component*

2.1 $\lim_{N\to\infty}\sup_t \Sigma_{k=-\infty}^{\infty}|E(\xi_t'\xi_{t+k}/N)| < \infty$ and
$\lim_{N\to\infty}\sup_{t,s} N^{-1}\Sigma_{i=1}^{N}\Sigma_{j=1}^{N}|cov(\xi_{is}\xi_{it},\xi_{js}\xi_{jt})| < \infty.$
2.2 The first eigenvalue of $E(\xi_t\xi_t')$, $\lambda_1^{\xi}$ is $O(N^{1-\alpha})$ with $0 < \alpha \leq 1$.

In our setting, the idiosyncratic error is serially and cross- correlated. The cross-correlation of $\xi_t$ is allowed to be strong among groups of periods but globally weaker than the cross-correlation of the common component. This is not the most standard assumption in economic applications, where $\xi_{it}$ are usually assumed to be weakly cross-correlated. However in our setting it is a more realistic assumption since $N$ increases with the sampling frequency, and the idiosyncratic components of the added periods will only be correlated with similar periods. The principal component estimator of the factors and the DI forecast is still consistent (see Heaton and Solo, 2006).

*Relation between the Idiosyncratic and Common Component*

3 $\xi_{it}$ and $\mathbf{F}_t$ are mutually uncorrelated for all $i,t$.

Under the preceding assumptions, the cross-covariance function of $\mathbf{y}_t$, $E(y_{it}y_{i-kt})$ which corresponds with the autocovariance is

$$E(y_{it}y_{i-kt}) = \Sigma_{j=1}^{r}\Lambda_{ji}\Lambda_{ji-k}\sigma_{jj} + E(\xi_{it}\xi_{i-kt}). \tag{7}$$

This shows that periodic correlation in our model has two sources, the heterogeneity of the loadings $\Lambda_{ji}$ and the heterogeneity across the idiosyncratic covariances $E(\xi_{it}\xi_{i-kt})$. Hence, the analysis of the loadings and idiosyncratic covariances is a useful tool to determine the presence of periodic correlation.

*Forecasting Equation*

4.1. $E(\mathbf{Z}_{ht}\mathbf{Z}_{ht}') = \Sigma_{Z_hZ_h}$ is a positive definite matrix with $\mathbf{Z}_{ht} = (\mathbf{F}_t, \mathbf{W}_{ht})'$.

4.2. $T^{-1}\Sigma_{t=1}^{T}\mathbf{Z}_{ht}\mathbf{Z}'_{ht} \xrightarrow{p} \Sigma_{Z_h Z_h}$.

4.3. $T^{-1}\Sigma_{t=1}^{T}\mathbf{Z}_{ht}\varepsilon_{ht+1} \xrightarrow{p} 0$.

4.4. $|\delta_h| < \infty$ with $\delta_h = (\Upsilon'_h, \beta'_h)'$ for all $i$.

Conditions 4.1-4.4 are required for the consistency of the least squares estimation of $\delta_h$.

# 3 Analysis of unit roots

So far it has been assumed that the $y_{it}$ variables (the hours $i$ within week $t$) have no unit roots. Given the large $T$ framework it needs to be checked formally. The presence of unit roots in $y_{it}$ can be due to the presence of unit roots in the factors or the presence of unit roots in the errors $\xi_{it}$. However, given that the (weekly) series correspond to the same underlying periodic (hourly) series, it is not likely that the idiosyncratic errors will contain unit roots. A similar argument applies that if there is a unit root this will be probably common among all hours, that is will be due to a common unit root factor.

Consider the case of one factor with a unit root, that is, $y_{it} = \Lambda_{1i}F_{1,t} + \xi_{it}$ where $\Delta F_{1,t} = u_t$ with $u_t$ is I(0). The presence of an I(1) factor implies that all $y_{it}$ are I(1) processes that are cointegrated with each other. The integration analysis could now be done separately for each $y_{it}$, but of course the unit root analysis for the common factor is a more powerful tool.

In the case the observed $y_{it}$ have a dynamic relation with the factors as in (4), the unit root analysis can be done with information criteria (see Bai, 2004). This method requires weak cross-correlated idiosyncratic components, an assumption that for some variables may not hold in the case the data are sampled at very high frequency. Hence, it is relevant to perform the unit root analysis for the raw data and for skip-sampled data.[2], see Stock and Watson (20020, Heaton and Solo (2006).

Continuing with the example when the loadings associated to the unit root factor are constant across $i$, $\Lambda_{1i} = \Lambda_1$, the periodic process $y_t$ is a I(1) process. In the case that factor loadings change with $i$, $y_t$ is a periodically integrated process (see Osborn, et al, 1988), and a quasi-difference operator is required to remove the unit root, that is, $y_t - \frac{\Lambda_{1i}}{\Lambda_{1i-1}} y_{t-1}$ is I(0) with $\frac{\Lambda_{1i}}{\Lambda_{1i-1}} \neq 1$. In case a unit root factor is found, the unit root can be removed from the periodic time series using the estimated quasi-difference operator, that is, $x_t = y_t - \frac{\widetilde{\Lambda}_{1i}}{\widetilde{\Lambda}_{1i-1}} y_{t-1}$, see Franses and Paap (2004).

# 4 Estimation

## 4.1 Common Factors

Under assumption 1.1 through 3 the common factors are estimated consistently with the method of asymptotic principal components (see Stock and Watson, 2002; Bai, 2003; Heaton and Solo, 2006). Under known $r$, the normalization $T^{-1}\Sigma_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t' = \mathbf{I}_r$ and concentrating out $\mathbf{\Lambda}$, $\widetilde{\mathbf{F}}_t$ is obtained as $\sqrt{T}$ times the eigenvectors of the $T \times T$ covariance matrix $\mathbf{yy}'$, associated to the $r$ largest sample eigenvalues, $\mu_1$,..., $\mu_r$, with $\mathbf{y}$ being a $T \times N$ matrix containing all information on $y_{it}$. The estimated loadings $\widetilde{\mathbf{\Lambda}}_i$ are obtained by regressing $\widetilde{\mathbf{F}}_t$ on $y_{it}$, and the common and idiosyncratic components are ob-

[2] In our illustration for hourly data, the unit root analysis is done for hourly data ($N = 168$), for data sampled every 2 hours ($N = 89$), every 3 hours ($N = 56$) and for every 4 hours ($N = 42$).

tained as $\widetilde{\chi}_{it} = \widetilde{\mathbf{\Lambda}}_i'\widetilde{\mathbf{F}}_t$ and $\widetilde{\xi}_{it} = y_{it} - \widetilde{\chi}_{it}$, respectively. The estimators $\widetilde{\mathbf{F}}_t$ and $\widetilde{\mathbf{\Lambda}}_i$ are conditionally Gaussian estimators of $\mathbf{RF}_t$ and $\mathbf{R}^{-1}\mathbf{\Lambda}_i$, respectively, where $\mathbf{R}$ is an $r \times r$ non-singular matrix.[3] This asymptotic distribution permits to construct confidence intervals and standard testing on for example the presence of periodic correlation due to heterogenous loadings.

The number of common factors $r$ is unknown and needs to be determined. Under weak cross-correlated errors, this can be done with information criteria (see Bai and Ng, 2002). Hence, given the probable presence of some strongly cross-correlated errors, it is convenient to check the robustness of the estimated $r$ to the sampling interval of the data, in a similar way as is done for the unit root analysis. It can be useful to examine the ratios of consecutive eigenvalues $\mu_{k+1}(N)/\mu_k(N)$ for data measured at different sampling intervals, as the $\mu_{r+1}(N)/\mu_k(N)$ is a consistent estimator of the lower bound for the noise to signal ratio $\lambda_1^{\xi}(N)/\lambda_r^X(N)$ where $\lambda_r^X$ is the $r$ largest eigenvalue of $E(\mathbf{y}_t\mathbf{y}_t')$ (see Heaton and Solo, 2006).

Given that the final aim of the factor analysis is forecasting, and, as will be discussed below, the final regressor set of the forecasting equation is chosen with a model selection method, it appears to be better to over-estimate the value of $r$ than to under-estimate it. The underlying factors are then still consistently estimated, and if some factors are not relevant for forecasting, the model selection will discard them. An under-estimation of $r$ will produce inefficient forecasts.

[3] The common component $\mathbf{\Lambda F}_t$ is observationally equivalent to $\mathbf{\Lambda R R}^{-1}\mathbf{F}_t$.

## 4.2 Forecasting Equation

To illustrate the validity of the DI model for forecasting, consider the case of a single factor model for an hours within a week variable, that is, $y_{it} = \lambda_i f_t + \xi_{it}$, where $f_t = \alpha f_{t-1} + u_t$ and $(1 - \Phi_i L^{168})\xi_{it} = \varepsilon_{it}$ with $|\Phi_i| < 1$. We can write

$$y_{it} = \alpha \left(\lambda_i - \Phi_i \lambda_{i-168}\right) f_{t-1} + \Phi_i y_{i-168t} + \eta_{it} \tag{8}$$

where $\eta_{it}$ is uncorrelated with both $f_{t-1}$ and $y_{i-168t}$. As seen from (8), in order for the factor to contain valuable information for forecasting $y_{it}$, the factor must be serially correlated ($\alpha \neq 0$) and also $\lambda_i \neq \Phi_i \lambda_{i-168}$ should hold.

Under known composition of $\mathbf{W}_{ht}$, the parameters of (5) are consistently estimated by least squares and denoted as $\widehat{\Upsilon}_h$ and $\widehat{\beta}_h$.

In practice, we do not observe the idiosyncratic component, and, in addition, the dynamics of the associated periodic time series $\xi_{it}$ are hard to specify. Therefore, we will use forecast equation selection methods to determine the specification of the idiosyncratic part of the forecasting equation. The usual practice in factor analysis is to use the BIC to determine the lag specification of $y_{ht-k}$ $k = 0, ..., K$. In our setting the set of regressors $\mathbf{W}_{ht}$ will also include lags of other periods than $h$, like for example $y_{h+1t}$, $y_{h-1t}$, $y_{h+168t}$, $y_{h-168t}$.

In applied research the most common selection method is the root mean square error computed in a simulated out-of-sample forecast. However, this method does not select consistently the best approximating model among the candidate models (see Inoue and Kilian (2006). Therefore, we propose to use the BIC criterion as it is a consistent selection method. An initial set of regressors $\mathbf{Z}^0_{ht}$ may include all the estimated factors and the series $y_{it}$ that are most cross-correlated in terms of idiosyncratic errors. The final forecasting equation specification $\mathbf{Z}^*_{ht} = (\mathbf{F}^*_{ht}, \mathbf{W}^*_{ht})$ minimizes BIC. Due to the presence of periodic correlation, we compare nested models obtained by discarding the least significant regressors from $\mathbf{Z}^0_{ht}$, and by continuing until $\mathbf{Z}^*_{ht}$ is found, see Inoue and Kilian (2006).

# 5 Diffusion Index Forecast

The serial correlation feature of the common factors and the heterogeneity of the factor loadings imply the presence of valuable forecasting information for all $N$ variables $y_{it}$. The presence of serial and cross-correlated errors requires to capture somehow such residual dynamics in the forecast as well. In the framework of factor type forecasting, a diffusion index (DI) forecast turns out to be the more flexible and robust way to do so in the presence of complex dynamics (see Boivin and Ng, 2005).

The estimated DI forecast of $y_{hT+1}$ is given by:[4]

$$\widehat{y}_{hT+1|T} = \widehat{\Upsilon}^{*\prime}_h \widetilde{\mathbf{F}}^*_{hT} + \widehat{\beta}^{*\prime}_h \mathbf{W}^*_{hT},$$

where $\widehat{\Upsilon}^*_h$ and $\widehat{\beta}^*_h$ are the least squares estimates of $\Upsilon^*_h$ and $\beta^*_h$, the associated coefficients to $\widetilde{\mathbf{F}}^*_{ht}$ and $\mathbf{W}^*_{hT}$.

The 95% prediction interval for $y_{hT+1}$ is

$$(\widehat{y}_{hT+1|T} - 1.96\sqrt{\widehat{\sigma}^2_{\varepsilon_h} + var(\widehat{y}_{hT+1|T})}, \widehat{y}_{hT+1|T} + 1.96\sqrt{\widehat{\sigma}^2_{\varepsilon_h} + var(\widehat{y}_{hT+1|T})}),$$

[4]When the sample ends in the middle of the week, the last period $N$ is associated with the period of the last available observation.

where $\widehat{\sigma}^2_{\varepsilon_h} = T^{-1}\Sigma_{t=1}^{T}\widehat{\varepsilon}_{ht+1}$ with $\widehat{\varepsilon}_{ht+1} = \widehat{y}_{ht+1|t} - y_{ht+1}$,

$$var(\widehat{y}_{hT+1|T}) = T^{-1}\widetilde{\mathbf{Z}}^{*\prime}_{hT}Avar(\widehat{\delta}^*_h)\widetilde{\mathbf{Z}}^*_{hT} + N^{-1}\widehat{\beta}^{*\prime}_h Avar(\widetilde{F}^*_{hT})\widehat{\beta}^*_h, \tag{9}$$

$$Avar(\widehat{\delta}^*_h) = (T^{-1}\Sigma_{t=1}^{T-1}\widetilde{\mathbf{Z}}^*_t\widetilde{\mathbf{Z}}^{*\prime}_t)^{-1}(T^{-1}\Sigma_{t=1}^{T-1}\widehat{\varepsilon}^{*2}_{ht+1}\widetilde{\mathbf{Z}}^*_t\widetilde{\mathbf{Z}}^{*\prime}_t)(T^{-1}\Sigma_{t=1}^{T-1}\widetilde{\mathbf{Z}}^*_t\widetilde{\mathbf{Z}}^{*\prime}_t)^{-1}, \tag{10}$$

$Avar(\widetilde{F}^*_T) = \widetilde{V}^{*-1}\widetilde{\Gamma}^*\widetilde{V}^{*-1}$, with $\widehat{\delta}^*_h = (\widehat{\Upsilon}^{*\prime}_h, \widehat{\beta}^{*\prime}_h)'$, $\widetilde{\mathbf{Z}}^*_T = (\widetilde{F}^{*\prime}_T, \mathbf{W}^{*\prime}_T)'$ and $\widetilde{V}^*$ a diagonal matrix consisting of the eigenvalues of $\mathbf{y}\mathbf{y}'/(TN)$ associated to $\widetilde{F}^*$, and where $\widetilde{\Gamma}^*$ is obtained from $\widetilde{\Gamma}$, a cross-section and heteroskedastic autocorrelation consistent estimator of $\Gamma$ (see Bai and Ng, 2006),

$$\widetilde{\Gamma} = n^{-1}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\widetilde{\mathbf{\Lambda}}_i\widetilde{\mathbf{\Lambda}}'_j T^{-1}\Sigma_{t=1}^{T}\widetilde{\xi}_{it}\widetilde{\xi}_{jt}, \tag{11}$$

where $n = \min(\sqrt{N}, \sqrt{T})$.[5]

# 6 Illustration

We compare the performance of the DI forecast with a naive method $\widehat{y}_{hT+1|T} = y_{hT}$ in a simulated real-time experiment with hourly load data from PJM Mid-Atlantic Region spanning the period January 7, 2002 to June 18, 2006.[6] Hence, $i$ denotes hour of the week and $N = 168$. No unit roots are detected with information criteria of Bai (2004) for the raw data nor for the various skip-sampled data. The information criterion of Bai and Ng (2002) suggests that there are 20 factors, but this amount seems not to be robust to skip-sampling. The inspection of the ratios of consecutive eigenvalues for different

[5]In our setting the cross section has a natural ordering and therefore HAC estimators of Newey and West (1987) or Andrews (1991) can be used instead of (11).

[6]The data were obtained from http://www.pjm.com.

sampled data suggests the presence of 6 factors.[7]

Figures 1 and 2 plot the estimated factors and loadings for the case of 231 weekly time series observations. The first factor seems to pick up a two cycles per week component that characterizes the weekly seasonality of electricity load. The remaining factors seem to capture cyclical features at a higher periodicity than the weekly component. The factor loadings present a clear periodic pattern, reflecting the periodic correlation feature of the underlying hourly load series. Each loading presents a particular intra-week and intra-day pattern.

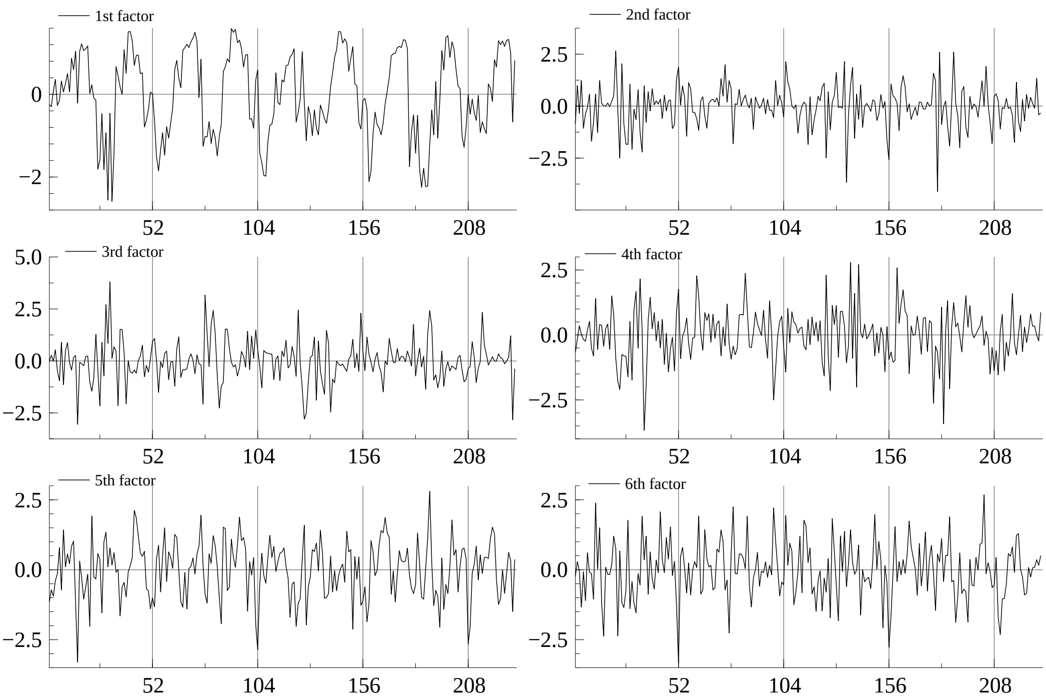
To examine the forecast performance, the above exercise is repeated. We start with the sample January 7-13, 2002 to November 15-21, 2004 (150 weeks) and we forecast $N$ hours ahead. The final set of regressors $\widetilde{\mathbf{Z}}^{*}_{hT}$ included in the forecast for each iteration and for each forecasting horizon is selected with BIC criteria.

Figure 3 plots the average MAPE for the horizons considered in the simulation, that is from one hour to one week ahead. The DI forecast model improves substantially the forecast performance of the seasonal random walk. We also plot the average MAPE corresponding to the same predicted period. The first point is the average of MAPEs of forecasting Monday 1 AM from all other hours of the week. As seen in Figure 4, some hours are easier to forecast than others, which again emphasizes the periodic correlation in the data.

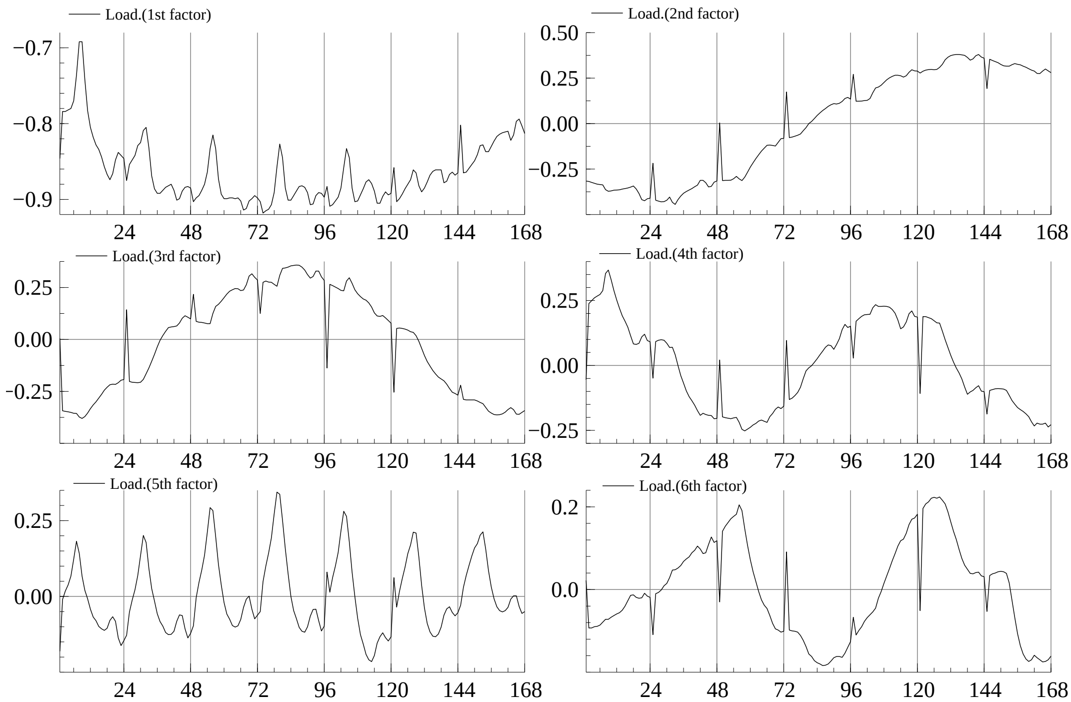
[7]We also performed an exercise assuming 10 factors, but we did not obtain substantially different forecasting performance.

# 7 Conclusion

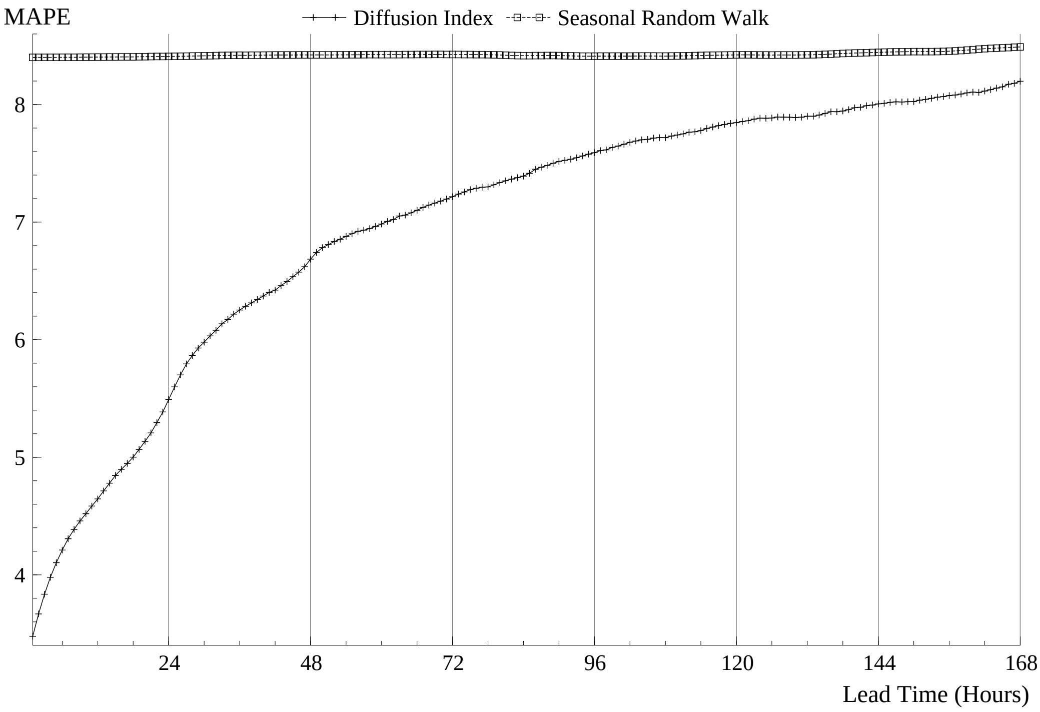
In this paper we proposed to use the so-called diffusion index model to describe and forecast high frequency data with periodic correlations. When there are many seasons within a cross section (here we have 168 hours within a week), standard periodic models would not work and hence a panel view on this issue seems most useful. We outlined representation, estimation, inference and forecasting, and we illustrated for hourly electricity data that our method works rather well.



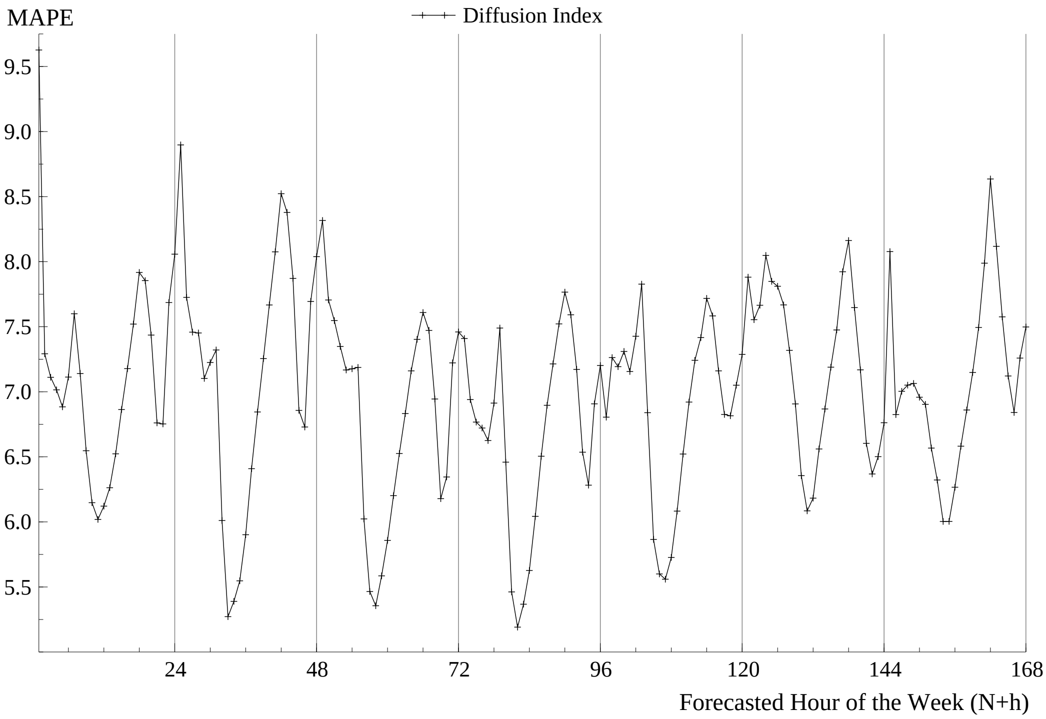

Estimated factors (January 7, 2002 to June 18, 2006)




Estimated loadings (January 7, 2002 to June 18, 2006)




MAPE results plotted against lead time




MAPE results plotted against predicted hour

# 8 References


Anderson, P.L. and A.V. Vecchia (1993) Asymptotic Results for Periodic Autoregressive Moving-Average Processes. *Journal of Time Series Analysis* 14, 1-18.

Andrews, D.W.K. (1991) Heteroskedastic and Autocorrelation Consistent Matrix Estimation. *Econometrica* 59, 817-854.

Bai, J. (2004) Estimating Cross-Section Common Stochastic Trends in Non-stationary Panel Data. *Journal of Econometrics* 122, 137-183.

Bai, J. and S. Ng (2002) Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70, 191-221.

____ (2004) A Panic Attack on Unit Roots and Cointegration. *Econometrica* 72, 1127-1177.

____ (2006) Confidence Intervals for Diffusion IndexForecasts and Inference for Factor-Augmented Regressions. *Econometrica* 74, 1133-1150.

Bessembinder, H. and M.L. Lemmon (2002) Equilibrium Pricing and Optimal Hedging in Electricity Forward Markets. *The Journal of Finance* 57, 1347-1382.

Boivin, J. and S. Ng (2005) Understanding and Comparing Factor-Based

Forecasts. *International Journal of Central Banking* 2, 117-151.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000) The Generalized Dynamic Factor Model: Identification and Estimation. *Review of Economics and Statistics* 82, 540-554.

Franses, P. H. (1994) A Multivariate Approach to Modelling Univariate Seasonal Time Series. *Journal of Econometrics* 63, 133-151.

Franses, P.H. and R. Paap (2004) *Periodic Time Series Models*, Oxford University Press, Oxford.

Hansen, L.P. and T.J. Sargent (1993) Seasonality and Approximation Errors in Rational Expectations Models. *Journal of Econometrics* 55, 21-55.

Harvey, A. and S.J. Koopman (1993) Forecasting Hourly Electricity Demand Using Time-Varying Splines. *Journal of the American Statistical Association* 88, 1228-1236.

Heaton, C. and Solo, V. (2006) Estimation of Approximate Factor Models: Is it Important to Have a Large Number of Variables? Paper presented at the North American Summer Meetings of the Econometric Society 2006, Minneapolis, Minnesota.

Inoue, A. and L. Kilian (2006) On the Selection of Forecasting Models. *Journal of Econometrics* 130, 273-306.

Lund, R., H. Hurd, P. Bloomfield, and R. Schmith (1995) Climatological Time Series with Periodic Correlation. *Journal of Climate* 8, 2787-2809.

Newey, W., and K. West (1987) A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55, 703-708.

Osborn, D.R., A.P.L. Chui, J.P. Smith and C.R. Birchenhall (1988) Seasonality and the Order of Integration for Consumption. *Oxford Bulletin of Economics and Statistics* 50, 361-377.

Pagano, M. (1978) On Periodic and Multiple Autoregressions. *Annals of Statistics* 6, 1310-1317.

Stock, J.H. and M.W. Watson (2002) Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association* 97, 1167-1179.

Tiao, G. C. and M. R. Grupe (1980) Hidden Periodic Autoregressive Moving Average Models in Time Series Data. *Biometrika* 67, 365-373.

Taylor, J.W., L.M. de Menezes and P.E. McSharry (2006) A Ccomparison of Univariate Methods for Forecasting Electricity Demand Up to a Day Ahead. *International Journal of Forecasting* 22, 1-16.