# Causal Analysis for Non-stationary Time Series in Sensor Rich Smart Buildings

Yuxun Zhou[1], Zhaoyi Kang[1], Lin Zhang[2], *Member, IEEE* and Costas Spanos[1], *Follow, IEEE*

***Abstract*—Advances in sensor rich smart buildings have given researchers access to a multitude of time series data, including temperature, humidity, real and reactive power consumption of specific nodes or devices, occupant presence and activities, etc. Time series generated by sensor networks reflect various phenomena in buildings and are naturally related to each other. Hence quantitative techniques are required to exploit dependence among different types of sequences in order to allow smart applications such as non-intrusive activity detection, energy usage prediction, demand side management and control. Past research of relational analysis has focused on symmetric correlative statistics. On the other hand, the asymmetric causal relation can capture more dynamic and complex relationships and is able to reveal directed influence among series. However, most traditional causal analysis relies on stationarity, while the statistics of real sensor measurement in smart buildings is rarely time invariant. In this paper, a statistical time series analysis framework is proposed to examine causal relationships among time series that are highly non-stationary. The Granger causality identification is extended to sensor data in buildings and the issue of non-stationarity is addressed firstly by using modified Hodrick-Prescott (HP) filter which is able to extract simpler trend components. Secondly, Autoregressive Integrated Moving Average model with exogenous variables (ARIMAX) model is trained for different components of two series. Finally, Granger causality is tested for both directions by F-statistics. The above procedure is performed on actual energy-consumption time series to exploit potential causal relations.***

## I. INTRODUCTION

With the wide deployment of power meters and sensor networks for smart buildings, researchers are now able to access large amount of real time sensory data, including the environment temperature, humidity in different areas, real and reactive power consumption of specific nodes or devices, etc. These measurements are time series streaming to data servers and providing information not only about power consumption, indoor environment in a straightforward way, but also device status, occupant activities, and many other measurable phenomena implicitly.

Proper information extraction by time series analysis is thus an indispensable enabler for applications in multiple fields such as energy saving, non-intrusive activity detection, smart power consumption management (demand response), building environmental monitoring, user comfort evaluation, and so on. Due to highly correlated underlying physical coherence, these measurements have inherent relations among themselves, which are also of great importance in providing required information or reasonable model for advanced smart applications. The past research on relational analysis focused on using symmetric techniques such as correlative statistics or mutual information, which are easy to compute but unable to determine directed dependence among related time series.

Causality is one of asymmetric measures that can capture this subtle directed influence and is able to reveal the fact one process statistically "causing" other processes. By time series causal analysis, more dynamic, complex relationships among different phenomena could be exploited, which could in turn benefit many upper level applications. For example, better control strategies could thus be designed by considering sequential influence among these processes. Also, causality has a close interplay with multivariate time series prediction in which proper selection of explanatory variables from a number of candidates is needed in order to reduce the problem complexity with minimal impact on the performance.

The idea of finding causality in time series can be dated back to the early 1960s, when Granger proposed his famous causality test by examining if one time series is useful in the prediction of another [1]. This notion of causality has been widely adopted, and has many applications in economic, industrial, biologic and sociological data analysis [2] [3]. Most recent works on causal analysis focus on finding a causal network with multivariate time series models [4], using bootstrap technique to enhance statistics [5], or analyzing causality by estimating directed information [6].

Most of the previous techniques require stationarity of time series. However, in the case of sensor in smart buildings, the time series generated are usually non-stationary and may possess daily and seasonal patterns. One possible way to get rid of non-stationarity is performing several orders of differencing [7], nonetheless, high order difference will lead to loss of causal information thus may not be able to reveal true causal relations.

In this paper, we address the non-stationary nature of our data streams by decomposing the original complex highly non-stationary time series into several simpler time series in different frequencies, and then perform extended Granger causality test for each of these components. The idea is that time series resulting from such decomposition will be much easier to model using fewer lags and differences, thus non-stationarity can be handled by common time series models and the power of statistical test could be maintained. On the other hand, it is reasonable to assume that causality occurs within certain sampling frequencies, thus the above

[1] Department of EECS, University of California at Berkeley, U.S.A. yxzhou@eecs.berkeley.edu, kangzy@eecs.berkeley.edu, spanos@eecs.berkeley.edu
[2] Department of Electronic Engineering, Tsinghua University, Beijing, China linzhang@tsinghua.edu.cn

decomposition is expected not to "obscure" causal relations. With these observations, firstly, a modified Hodrick-Prescott (HP) filter is applied to extract trend components from the time series in question, and then an ARIMAX model is trained for the resulted components, and finally Granger causality is tested for both directions by F-statistics.

The content is arranged as follows: in Section II we briefly describe the non-stationary nature and some other characteristics of smart building sensor measurements which would inhibit previous causal analysis methods. In Section III, a modified Hodrick-Prescott filter is formulated to decompose the original time series into different components. Section IV is devoted to causality identification and testing based on the ARIMAX model, where we give a succinct introduction to the time series model and describe a statistical testing method for the verification of Granger causality. Section V presents the case study of the proposed algorithm on two real-world datasets. Finally, Section VI concludes the paper.

## II. SMART BUILDING ELECTRICITY METERING AND SENSOR MEASUREMENT

In typical smart buildings, meters are placed at critical nodes of the electrical tree of the building and provide high resolution readings of real/reactive power consumption in each phase, voltage, power factor, and accumulated energy usage, etc. Figure 1 shows a Dent™ meter readings once a second and plotted over a 24 hour interval.

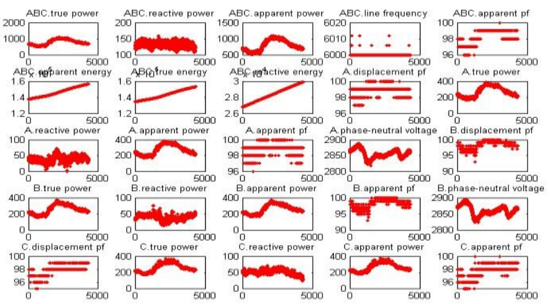


Figure 1. Cory Hall power meter readings [14].

Besides smart electricity meters, many other types of sensors are now deployed in buildings to monitor environmental processes such as air quality, humidity, temperature, CO2 concentration, etc. Figure 2 shows an indoor particulate matter (PM) sensor reading for two days in the Berkeley CREST Center. It is seen that 24-hour plots show fluctuations as well as diurnal (daily) patterns, for example, the PM sensor data in Figure 2 exhibit daily maximum during daytime and minimum in the night. Moreover, for long-term sensor data such as temperature for a whole year, yearly (seasonal) trend is combined with weekly and daily patterns. As is mentioned before, a major challenge for causality analysis with this kind of time series which have seasonal trends is that the stationary assumption does not apply and simple time series models such as multivariate regression might generate auto correlated errors (thus incorrect test statistics). On the other hand, complex models like Seasonal ARIMA with exogenous variables (SARIMAX) are hard to train, and introducing too many training coefficients reduces the statistical power for causality testing.

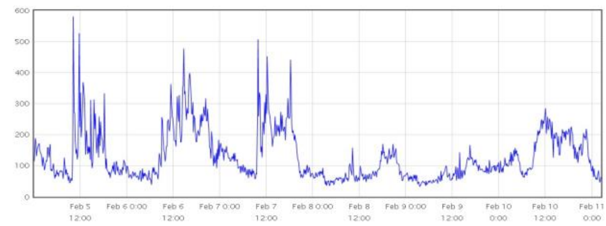

Figure 2. PM readings in CREST Center.

If we consider in frequency domain two time series on which we would like to perform causal analysis, intuitively, causal influence should occur within comparable time constants for the two series. For example, PM series' fluctuation in time resolution of minutes should be caused by occupant motions in similar timing, while daily or monthly changes of PM measurement should be explained better by exogenous variables with slower trends such as weather conditions or outdoor air quality. Inspired by the fact that causal relation between two phenomena usually reveals itself in similar regions in the frequency spectrum we propose the following time series decomposition method that paves the way for further causality identification.

## III. TIME SERIES DECOMPOSITION USING MODIFIED HP FILTER

Causality, by definition, has a notion of "sequencing in time", thus for our purpose a filter is required to do efficient decomposition while maintaining timing information and sequential causal relations. Actually, many decomposition filters are proposed in time series analysis and are quite different in complexity, accuracy and selectivity. The Hodrick-Prescott (HP) filter suggested by Robert J. Hodrick and Edward C. Prescott [8] is a promising one for causal analysis in that it works efficiently in time domain. Further, it is flexible in tuning out different components by simply changing smoothness parameters, which is more straightforward compared with other time series decomposition methods.

### A. *Modified Recursive HP filter formulation*

The original version of HP filter could be found in [8]. It is widely applied in economics in order to extract long-term trends in time series. In this paper, we modify the HP filter into a recursive version which is able to extract multiple sequential trends. The idea is to recursively subtract estimated longer term components and apply filtering process again to residuals so as to obtain components with higher frequency. Let the original time series be expressed by a sum of components which represent fluctuations in different frequencies:

$$x_t = x_t^1 + x_t^2 + \ldots + x_t^n \tag{1}$$

For example, $x_t^1$ could be monthly trend of original time series, $x_t^2$ and $x_t^3$ could be weekly and daily trends, respectively. Then the decomposition procedure is recursively formulated as solving the following unconstrained optimization:

$$x_t^{i+1} = \arg\min_{\{u_t\}} \sum_t \left( x_t - \sum_{k=1}^{i} x_t^k - u_t \right)^2 + \lambda_{i+1} \sum_t \left[ (u_{t+1} - u_t) - (u_t - u_{t-1}) \right]^2 \quad (2)$$

Note that in the beginning, we estimate $x_t^1$ by solving

$$x_t^1 = \arg\min_{\{u_t\}} \sum_t \left( x_t - u_t \right)^2 + \lambda_1 \sum_t \left[ (u_{t+1} - u_t) - (u_t - u_{t-1}) \right]^2 \quad (3)$$

### B. *Interpretation and Parameter specification*

Intuitively, the first term in Eq. (2) penalizes the squared error between the current component under estimation and the time series residual from the previous step. The second term penalizes the second order derivative of the component so that the extracted component is expected to be "smooth", with the level of "smoothness" determined by the value of seasonal $\lambda_i$. Higher value of $\lambda_i$ indicates more emphasis on the smoothness of the current trend component, which results in slower variant trend components with longer time scale. The optimization problem involved in the HP filter admits explicit solution which makes it easy to implement. The only trick here is the choice of proper smoothness weights $\lambda_i$ such that the component of the specific frequency that we are interested in can be extracted. There are many works in the literature that address the choice of "smoothness" parameter in HP filter, however, currently no automatic selection algorithm is available and empirical values are usually used. In [9] Ravn and Uhlig proposed to pick lambda as the fourth power of the frequency observation ratio, which is 6.25 for annual data 1600 for quarterly data, and 129,600 for monthly data.

## IV. Causality analysis based on ARIMAX and F-testing

Given two time series $x_t$ and $y_t$, the existence of Granger causality from $y_t$ to $x_t$ is defined by the fact that past values of $y_t$ can provide statistically significant information about future values of $x_t$. The main idea of the Granger causality test comes from the definition itself: build tentative prediction models with both time series under consideration, and then test the statistical hypothesis that would verify the fitness of the model. In this section, firstly a simple version of Granger causality test using F-statistic is introduced, then we extend the basic idea but use advanced time series model which is able to handle reduced non-stationarity of the time series generated from HP filter.

### A. *Original form of Granger causality analysis*

If time series $x_t$ and $y_t$ are stationary, a simple Auto Regression model (AR) could be used:

$$y_i = \sum_{k=1}^{p} \alpha_k y_{i-k} + v_i \quad (4)$$

Also, for combined $x_t$ and $y_t$, an ARX model can be fitted to forecast $y_t$:

$$y_i = \sum_{k=1}^{p} [\alpha_k y_{i-k} + \beta_k x_{i-k}] + w_i \quad (5)$$

The order of the model is chosen according to some information criterion. From the notion of Granger causality, lagged values of $x$ are retained in model (5) if they add significant explanatory power to the regression. Thus the existence of causality would imply that the residuals in (5) are significantly smaller than that of model (4). From this viewpoint an F-test can be designed for statistics of $v_i$ and $w_i$.

### B. *ARIMAX time series models*

As is mentioned before, the preceding procedure does not apply for non-stationary seasonal data. While in section III by using HP filter we are able to decompose the original complex time series into several simpler components, the resulting components time series may not be "good" enough for simple regression models. They may still possess non stationary terms but may be suitable for advanced time series models in their simpler forms since long-term seasonality has been eliminated.

The Autoregressive Integrated Moving Average (ARIMA) model [10] is one of the milestones in linear time series analysis. It can be considered as a combination of two basic operations: lagging and differencing. As is implied by its name, the model contains three parts, auto regression of order $p$, integration of order $d$, and moving average of order $q$. An extension of ARIMA model is to incorporate exogenous variables (or called explanatory variables) into the prediction of future output values. This scheme is usually referred to as ARIMAX model, which can be summarized as follows: suppose we would like to model time series $x_t$ with another time series $y_t$ as the exogenous variable, then ARIMAX model contains:

$$z_t = \nabla^{d_1} x_t \quad (6)$$

$$w_t = \nabla^{d_2} y_t \quad (7)$$

$$(1 - \sum_{k=1}^{p} \alpha_k L^k) z_t = \mu + \sum_{k=1}^{q} \beta_k L^k w_t + (1 + \sum_{k=1}^{r} \gamma_k L^k) e_t \quad (8)$$

where $\nabla$ is the difference operator, and

$$\nabla x_t = x_t - x_{t-1} \qquad \nabla^d x_t = \underbrace{\nabla \cdots \nabla}_{d}(x_t) \quad (9)$$

$L$ is the lagging operator,

$$L x_t = x_{t-1} \qquad L^k x_t = x_{t-k} \quad (10)$$

and $\alpha_k$, $\beta_k$, $\gamma_k$ are model parameters. Equations (6) and (7) are the integration part that ensures stationary series in (8) by differencing $x_t$ and $y_t$, $d_1$ and $d_2$ times, respectively. The left hand side in equation (8) is the auto regression part, which basically describes dependence of present observations on its own past values. $\mu$ is the non-zero mean difference between time series $z_t$ and $w_t$ . The second term on the right hand side of equation is the contribution of past values of $w_t$ to current readings of $z_t$ , thus the impact of a finite memory of exogenous variables on predictions of $x_t$ is taken into account. The last term in the right hand side of (8) is the moving average part, which accounts for the correlation of current measure of $z_t$ and previous prediction errors.

### *C. F-Test for Granger causal analysis*

From the above ARIMAX model we are able to perform Granger causality test by directly adopting the notion. Since the influence of exogenous variable is totally captured by the second term in (8), the null hypothesis that there is no causal influence from time series $y_t$ to $x_t$ should be

$$H_0: \quad \beta_k = 0 \quad \forall k \tag{11}$$

If the null hypothesis is accepted, then it means that a model that only considers the past information of output variable is as good as the one incorporates information of exogenous variable. Thus when null hypothesis is verified, a simpler ARIMA model could be formulated for $x_t$

$$z_t = \nabla^d x_t \tag{12}$$

$$(1 - \sum_{k=1}^{p} \alpha_k L^k) z_t = \mu + (1 + \sum_{k=1}^{r} \gamma_k L^k) e_t \tag{13}$$

And this model is expected to have comparable performance as the model described by Eqs. (6) - (10). On the contrary, if the null hypothesis is rejected, then the prediction of $x_t$ can be significantly improved by adding past values of $y_t$, which in turn establishes the causal relation.

Now it is possible to compare residuals to test the "goodness of fit" of the two alternative models (8) and (13), and this is the main idea of the following statistical test. In fact, the hypothesis testing problem discussed before is equivalent to an F-test for the residuals of the two models. Let $RSS_1$ and $RSS_2$ denote the residuals of model (13) and (8), respectively, which can be calculated as

$$RSS = \sum_{t=1}^{T} e_t^2 \tag{14}$$

For the two candidate models, the statistic $s$ follows an F distribution:

$$s = \frac{(RSS_1 - RSS_2)/q}{RSS_2 / [T - 2(p+q+r) - 1]} \sim F_{q, T-2(p+q+r)-1} \tag{15}$$

If s is larger than a specified critical value of the corresponding F distribution, the null hypothesis is rejected and we can say that $y_t$ has a significant causal influence on $x_t$.

Note that by this type of hypothesis test, only the significant existence of causality is approved. Thus the above hypothesis testing is prone to reject weak causality and lacks straightforward interpretation. A measure inspired by information theory is proposed in [11] to indicate comparative causality "strength":

$$g(y \to x) = \log \frac{RSS_1}{RSS_2} \tag{16}$$

In the following experiments, we calculate both the $s$ statistic and indicator $g$ to exploit causal relations.

## V. CASE SPECIFICATION AND RESULTS

The aforementioned causal analysis scheme is performed on two data sets which are collected by smart meters and sensors installed at the University of California, Berkeley (UCB) and the Laurence Berkeley National Laboratory (LBNL) facilities. One is LBNL Building 90 (B90) data [13] for one year's temperature and electricity consumption measurement. The other is three months' high resolution power meter reading in Cory Hall at UCB [14]. In the first experiment, we show the proposed decomposition and testing procedure for the two time series. In the second experiment, in which multiple time series are involved, we perform an exhaustive search among all possible pairs of time series to build a multi-layer causal network which implies causal relations among power consumptions in different locations.

### *A. Causal analysis with B90 data*

As a first test for the proposed decomposition and corresponding time series model, this data set measures hourly average electricity consumption for the whole building in one year and average temperature with the same length and time resolution. Through a modified HP filter, we extract three components of monthly, weekly, and daily trends. The important smoothness parameter is chosen according to the discussions in section III. Note that some "repair" work has to be done in advance to replace missing values or most obvious sensor errors by their neighboring readings. Figures 3 and 4 show the decomposition results. In each figure, the left top is the repaired original signal, and the three sub figures present the decomposed monthly, weekly, and daily components. The results indicate that the proposed modified HP filter can efficiently find components in different frequencies.

As is discussed in section IV, an ARIMAX model is built for two series with one of them considered as exogenous variable, and an ARIMA model is also built for the other time series. The order of the model, i.e. $p$, $q$, $r$ are selected according to Akaike information criterion [12] using a package in R. Then residuals of the two models are obtained by one step prediction of the model on the same data set.

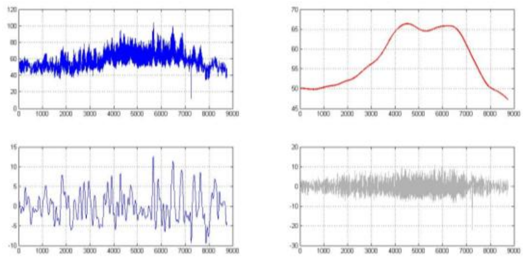

Figure 3. Decomposition for temperature time series of B90. $x$ axis: time evoluiton with one hour resolution; $y$ axis: temperature ( ºF) Top left: original measurement; Top right: monthly component; Bottom left: weekly component; Bottom right: daily component

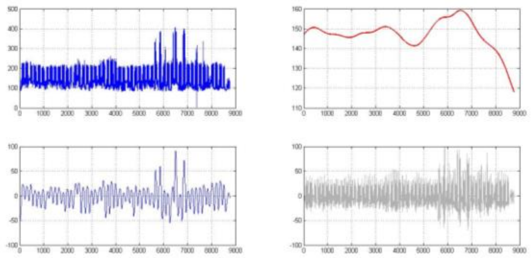

Figure 4. Decomposition for electricity consumption time series of B90. *x* axis: time evolution with one hour resolution; *y* axis: real power ( Kwatt) Top left: original measurement; Top right: monthly component; Bottom left: weekly component; Bottom right: daily component

Finally, residual ratios and F-test critical values are computed for hypothesis testing. Table I shows the results. We see that temperature has a strong weekly causal influence on electricity consumption, while the daily causal relation is much weaker than weekly relations. No significant monthly causality is found from temperature to electricity consumption since the null hypothesis is accepted. Also, the Granger ratio defined in section 4 is calculated as an indicator for the strength of causality. Similarly, causal analysis for the inverse direction can be performed. Not surprisingly, it is found that causal relation from electricity consumption to temperature is weaker than values that in the reverse direction.

TABLE I. CAUSAL ANALYSIS FROM TEMPERATURE TO ELECTRICITY CONSUMPTION

| x: temperature y: electricity | **0.95 F critical** = 2.80 | | | |
|---|---|---|---|---|
| | **Value y to x** | | **Value x to y** | |
| | ***s Statistic value*** | ***Granger ratio*** | ***s Statistic value*** | ***Granger ratio*** |
| ***Monthly*** | 0.766 | 3.7e-04 | 2.63 | 0.001 |
| ***Weekly*** | 173.3 | 0.083 | 211.8 | 0.101 |
| ***Daily*** | 45.8 | 0.023 | 59.2 | 0.029 |

## B. *Causal analysis with Cory Hall data*

As a second test, multiple time series with a higher time resolution are considered. We collect Dent$^{TM}$ meter measurements of 18 nodes in different locations of the electric grid of Cory Hall at UC Berkeley [14]. Meter readings are stored in a server every 20 seconds including 3 phase real and reactive power, apparent power, power factor, accumulated energy, etc. In Table II, we select several typical meters and causal analysis is performed over the one month data (of April 2011). Again, some repair work is done to take care of missing values and obvious false readings.

In Figure 5 and 6, decomposition results are shown for "lighting transformer" (1A) and "east lab" (1C) time series. The smoothness parameter is taken to be the fourth power of the frequency ratio with respect to monthly value. Thus the three consecutive sub figures are approximately the weekly, daily, and hourly components.

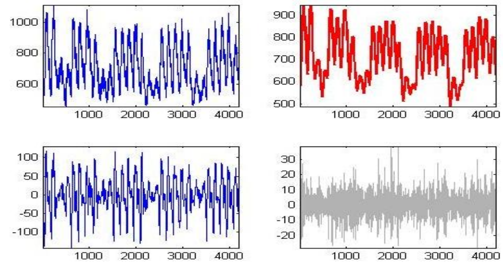


Figure 5. Decomposition for time series 1A. *x* axis: time evolution with 20 seconds resolution; *y* axis: apparent power consumption( KVA) Top left: original measurement; Top right: weekly component; Bottom left: daily component; Bottom right: hourly component

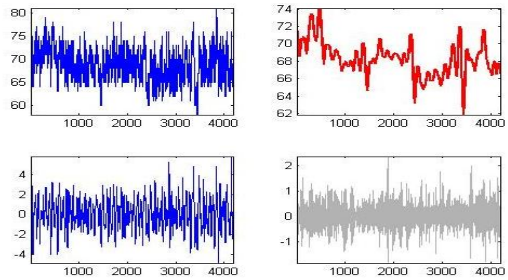


Figure 6. Decomposition for time series 1C. *x* axis: time evolution with 20 seconds hour resolution; *y* axis: apparent power consumption( KVA) Top left: original measurement; Top right: weekly component; Bottom left: daily component; Bottom right: hourly component

As previously, ARIMAX and ARIMA models are trained in each causal test. In table III to VI, s statistics and Granger causality indicator for each pairs of hourly and daily component are shown. Weekly analysis is ignored because one month data is far from enough to draw conclusions about weekly phenomena. Note that in all tables, causality is tested from component in the row to corresponding component in the column. And F critical value is found to be 3.79 at the 95% confidence level.

TABLE II. SELECTED SMART METERS IN CORY HALL

| **Meter NO.** | **Description** |
|---|---|
| ***1A*** | Lighting transformer |
| ***1C*** | 1st Floor East Lab |
| ***2D*** | East Power riser |
| ***2E*** | Fifth floor Plug and lighting loads |
| ***2F*** | Switch block 10 structure lighting |
| ***3C*** | Switch block for feeding HVAC and elevators |

Some interesting observations can be made. Firstly, the estimated causal relations are non-symmetric, as expected. For example, in hourly analysis in Table III, a significant causality can be found from 1A to 2E, while in the reverse direction, 2E to 1A, the existence of causality is rejected by F test. Secondly, causal dependence behaves differently in different frequencies. For example, hourly speaking, the East Power riser, namely 2D, does not causally influence any other measurement. While in the daily analysis, it shows a strong causal impact on others. In addition, some measurement always has a great causal influence on another. For example, it is found that Lighting Transformer (1A) has a significant causal relation to the fifth floor plug and lighting loads (2E), both in hourly or daily analysis. It may correspond to the fact that 1A is measuring total lighting loads while 2E represent some parts of it. It is worth well to point out that negative values in table III and V means that ARIMAX model cannot give better prediction than ARIMA model, because it requires more parameters to fit and the exogenous variable does not contain useful information.

With the resulted s statistics and g indicator, we are able to draw conclusions about causal relations among different sensor measurements, which reflect causal relations of electricity usage for activities in the building. Thus our causal analysis can provide useful information for building activity modeling, energy consumption scheduling, event control, etc.

TABLE III. S STATISTICS FOR EACH PAIRS (HOURLY COMPONENT)

| S | 1A | 1C | 2D | 2E | 2F | 3C |
|---|---|---|---|---|---|---|
| **1A** | / | -42.21 | 9.129 | 3.075 | 21.43 | 13.69 |
| **1C** | 8.832 | / | 23.15 | 20.81 | 10.18 | 1.698 |
| **2D** | 0.6840 | -37.77 | / | 6.446 | 27.52 | 5.400 |
| **2E** | 41.167 | 8.5236 | 0.1098 | / | 1.741 | 2.226 |
| **2F** | 20.71 | 23.34 | 22.03 | 22.91 | / | 6.771 |
| **3C** | 5.197 | -54.56 | 4.667 | 3.113 | 5.854 | / |

TABLE IV. G VALUE FOR EACH PAIRS (HOURLY COMPONENT)

| G | 1A | 1C | 2D | 2E | 2F | 3C |
|---|---|---|---|---|---|---|
| **1A** | / | -0.044 | 0.0094 | 0.0031 | 0.0221 | 0.0141 |
| **1C** | 0.009 | / | 0.0238 | 0.0214 | 0.0105 | 0.0017 |
| **2D** | 0.0007 | -0.0398 | / | 0.0066 | 0.0282 | 0.0055 |
| **2E** | 0.0421 | 0.0088 | 0.0001 | / | 0.0018 | 0.002 |
| **2F** | 0.0213 | 0.0240 | 0.0226 | 0.0236 | / | 0.0070 |
| **3C** | 0.0053 | -0.0570 | 0.0048 | 0.003 | 0.0060 | / |

TABLE V. S STATISTICS FOR EACH PAIRS (DAILY COMPONENT)

| S | 1A | 1C | 2D | 2E | 2F | 3C |
|---|---|---|---|---|---|---|
| **1A** | / | 3.0213 | 0.7143 | -49.27 | 11.04 | 0.8606 |
| **1C** | 1.447 | / | 2.902 | -50.26 | 39.48 | 3.864 |
| **2D** | 93.44 | 4.3860 | / | -47.71 | 57.77 | 0.3174 |
| **2E** | 235.7 | 2.2128 | -0.3368 | / | 34.33 | 8.794 |
| **2F** | 4.779 | 5.0504 | 0.1031 | -99.90 | 0 | 2.111 |
| **3C** | 0.4093 | 1.0019 | 1.993 | -100.8 | 14.89 | / |

TABLE VI. G VALUE FOR EACH PAIRS (DAILY COMPONENT)

| G | 1A | 1C | 2D | 2E | 2F | 3C |
|---|---|---|---|---|---|---|
| **1A** | / | 0.0031 | 0.0007 | -0.0521 | 0.0114 | 0.0008 |
| **1C** | 0.0015 | / | 0.0030 | -0.0532 | 0.0404 | 0.0040 |
| **2D** | 0.0939 | 0.0045 | / | -0.0504 | 0.0587 | 0.0003 |
| **2E** | 0.2262 | 0.0022 | -0.0003 | / | 0.0352 | 0.0091 |
| **2F** | 0.0049 | 0.0052 | 0.0001 | -0.1076 | / | 0.0022 |
| **3C** | 0.0004 | 0.0010 | 0.0020 | -0.1087 | 0.0153 | / |

### *C. Forecast results for Cory Hall data*

From the previous part, causal relations are obtained for six time series of power consumption. As a direct application of causality, we are able to select exogenous variables to help forecasting future values of another time sequence. For example, it is found that in April, time series 2F has a significant causal influence on 1C in both hourly and daily analysis. Since this causality may reveal relations of user activities in different floors, we can assume that it is persistent not specifically in April but also in other months. The following results show that including time series 2F is beneficial in the forecasting of 1C for the data of May 2010. Figure 7 shows the one step forecasting results of time series 2F with (top sub-figure) or without (bottom sub-figure) exogenous variable 1C. Also, the Mean Squared Error is found to be 5.173 and 5.931 with and without the explanatory variable, respectively

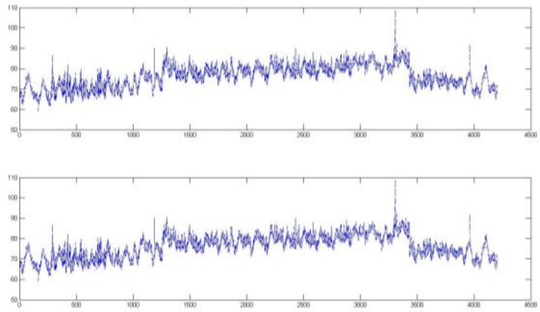

Figure 7. Forecasting reults for time series 1C. x axis: time evolution with 20 seconds reslution; y axis: Apparent power (KVA). Top: forecasting with 2F as exogenous variable; Bottom: forecasting using past values of 1C alone

## VI. CONCLUSION

In this paper, based on modified HP filter and ARIMAX model, we proposed a series analysis framework to identify causal relationships among senor measurements. Results show that the modified Hodrick-Prescott (HP) filter can efficiently extract trend components on different time scales with a specified smoothness parameter. Thus the original complex time series are decomposed into several simpler components, which could be well fitted by ARIMAX model. Granger causality is tested by using F-statistics for residuals of the time series model and experiment on real building data shows that causality is quite different from symmetric relational measures, and is able to reveal more dynamic and complex dependence among time series.

Our current work consider only pairwise causalities, in real life, however, causal dependence might be intertwined within multiple phenomena. Thus multi-variable models should be applied to examine cross causal dependence. Moreover, the proposed framework is only applicable for continuous time series. For discrete cases, such as measurement of user activities, directed information based causal analysis might be a potential alternative.

## ACKNOWLEDGMENT

This research is funded by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore

## VII. .REFERENCES

[1] Granger, C. W. J "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". *Econometrica* 37 (3): 424–438

[2] W.-K. Friston, K., Harrison, L., & Penny, W. "Dynamic causal Modeling" *NeuroImage*, 2003, 19(4), 1273–1302.

[3] Hesse, W., Möller, E., Arnold, M., & Schack, B. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124(1), 27–44.

[4] Dhamala, M., Rangarajan, G., & Ding, M. "Analyzing information flow in brain networks with nonparametric Granger causality". *NeuroImage*, 2008, 41(2), 354–362.

[5] Hatemi-J, A. and Roca, E. "An examination of the equity market price linkage between Australia and the European Union using leveraged bootstrap method", *European Journal of Finance*, 2008, 10, 475–88..

[6] Al-khassaweneh, M., & Aviyente, S. "The relationship between two directed information measures" *IEEE Signal Processing Letters*, 2010, 15, 801–804.

[7] John R. FreemanReviewed "Granger Causality and the Times Series Analysis of Political Relationships" *American Journal of Political Science*, Vol. 27, No. 2 May, 1983, pp. 327-358

[8] Hodrick, Robert, and Edward C. Prescott, "Postwar U.S. Business Cycles: An Empirical Investigation," Journal of Money, Credit, and Banking, 1997, 29 (1), 1–16.

[9] Morten O. Ravn, Harald Uhlig,"On Adjusting the HP-Filter for the Frequency of Observations" *The Review of Economics and Statistics*, Vol. 84, No. 2 (May, 2002), pp. 371-376

[10] Mills, Terence C. *Time Series Techniques for Economists*. Cambridge University Press, 1990.

[11] Barron, A., & Cover, T. "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, 1991, 37(4), 1034–1054.

[12] Brockwell, Peter J.; Davis, Richard A. *Time Series: Theory and Methods (2nd edition)* Springer

[13] Richard Brown, Steven Lanzisera, Hoi Ying (Iris) Cheung, Judy Lai, Xiaofan Jiang, Stephen Dawson-Haggerty, Jay Taneja, Jorge Ortiz, and David Culler. "Using Wireless Power Meters to Measure Energy Use of Miscellaneous and Electric Devices in Buildings." *6th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL 2011)*, Copenhagen, Denmark. 2011.

[14] Randy H. Katz, David E. Culler, Seth Sanders, Sara Alspaugh, Yanpei Chen, Stephen Dawson-Haggerty, Prabal Dutta, Mike He, Xiaofan Jiang, Laura Keys, Andrew Krioukov, Ken Lutz, Jorge Ortiz, Prashanth Mohan, Evan Reutzel, Jay Taneja, Jeff Hsu, Sushant Shankar. "An Information-Centric Energy Infrastructure: the Berkeley View" *Sustainable Computing: Informatics and Systems* 1 2011 7-22.