Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 145 (2016) 1059 – 1065

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Sustainable Design, Engineering and Construction

A Non-Stationary analysis using Ensemble Empirical Mode Decomposition to detect anomalies in building energy consumption

Hariharan Naganathan[a], Wai K. Chong[a]*, Zigang Huang[b], Ying Cheng [c]

[a] *Graduate student, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85287, United States*

[a*] *Associate Professor, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85287, United States*

[b] *Research Scientist, Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287, United States*

[c] *Chair Professor, Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287, United States*

Abstract

Commercial buildings' consumption is driven by multiple factors that include occupancy, system and equipment efficiency, thermal heat transfer, equipment plug loads, maintenance and operational procedures, and outdoor and indoor temperatures. A modern building energy system can be viewed as a complex dynamical system that is interconnected and influenced by external and internal factors. Modern large scale sensor measures some physical signals to monitor real-time system behaviors. Such data has the potentials to detect anomalies, identify consumption patterns, and analyze peak loads. The paper proposes a novel method to detect hidden anomalies in commercial building energy consumption system. The framework is based on Hilbert-Huang transform and instantaneous frequency analysis. The objectives are to develop an automated data pre-processing system that can detect anomalies and provide solutions with real-time consumption database using Ensemble Empirical Mode Decomposition(EEMD) method. The finding of this paper will also include the comparisons of Empirical mode decomposition and Ensemble empirical mode decomposition of three important type of institutional buildings.

Keywords: Empirical mode decomposition; Anomaly Detection; Commercial building; Hilbert Transform; Supply-Demand Characteristics

* Corresponding author. *E-mail address:* ochong@asu.edu

doi:10.1016/j.proeng.2016.04.137

1. Motivation

A modern building energy system can be viewed as a complex dynamical system that is interconnected and influenced by external (weather) and internal (system efficiency) factors. Modern large scale sensor and tracking devices can be deployed to measure some physical signals to monitor real-time system behaviors. These devices generate dynamic, diverse and large dataset and signals that have the potential to transform the management of buildings. Such data has the potentials to detect anomalies, identify consumption patterns, determine supply-demand characteristics, and analyze peak loads. The project team plans to develop a generic and systematic framework to detect hidden anomalous dynamical events, pre-process and analyze the data, and then process the analysis to aid the design and management of building energy design and systems. The mathematical foundation of the proposed framework is the Hilbert-Huang transform and instantaneous frequency analysis. The reason for this choice lies in the recognition that complex infrastructure systems are non-linear and non-stationary. For such systems, the traditional Fourier and wavelet transform based analyzes are limited because, fundamentally, they are designed for linear and stationary systems. The Hilbert-Huang transform and instantaneous frequency based analysis have proven to be especially suited for data from complex, non-linear, and non-stationary dynamical systems [1], [2].

2. Background

Post-Occupancy Evaluations (POE) [3], [4], data-mining [5], model calibration [6]–[8], statistical analysis [9], [10], and investment analysis [11], [12] were commonly used to narrow the gaps between designs and operations in building energy modelling. However, these methods do not accurately generate sufficient information to connect existing design and operational performances[13]. Energy design involves connecting the intimate lifecycle relationships between energy demand and supply, and the successful connection would propel energy efficiency to the next level where energy losses would be accurately estimated, and integrated into energy design. The feedback from operation and factors identification is critical in closing the design-operation gap [13]–[19]. The level of relationships between factors and time vary, for example, occupancy rate depends heavily on time while humidity does not. These factors, however, affect energy system performances indirectly and directly.

Traditional methods such as the Fourier transform and wavelet analysis assume stationarity and approximate the physical phenomena with linear models. These approximations may lead to spurious components in their time-frequency distribution diagrams if the underlying signal is non-stationary and nonlinear. The authors of previous literature that the Empirical Mode Decomposition (EMD) is a technique [20] to deal specifically with non-stationary and nonlinear signals. Given such a signal, EMD decomposes it into distinct modes, the intrinsic mode functions (IMFs), each having a distinct time or frequency scale and preserving the amplitude of the oscillations in the frequency range. The decomposed modes are orthogonal to each other, and the sum of all modes gives the original data. The ease and accuracy with which one uses the EMD method to process non-stationary and nonlinear signals have led to its widespread use in various applications such as seismic data analysis [1], and chaotic systems analysis [2], [21], neural signal processing in biomedical science and engineering [22], meteorological data analysis [9], and image processing [23].

While these methods are successful in identifying anomalies, the process of EMD has difficulties of high oscillations during reiterating IMFs at a different level. To overcome this issue, the paper proposes Ensemble Empirical mode decomposition (EEMD) algorithms that flatten the variation of oscillation during data discontinuities. Also, the methodology has a greater accuracy of results with energy data when compared to EMD.

3. Objectives and Methodology

The objective is to develop an automated computational method to detect, characterize, and understand anomalous dynamical behaviors from big energy data sets. The steps include: (1) perform EEMD and calculate distinct IMFs, (2) determine anomalies based on the amplitudes of IMFs, and (3) classify the anomalies regarding their frequencies. Unlike Fourier transform that usually becomes ineffective in a time-series analysis when the signal frequency changes with time, EEMD is well suited for generating IMFs where frequencies vary with time when the IMF period is a function of time: $T = T(t)$.

3.1 Data Pre-processing

In reality, it is not possible to preserve the integrity of the large and complex data set, especially for continuous recording over a long period, during which disturbances in the experiments or malfunctioning of sensors and detectors are inevitable. It is typical that most files would contain various segments of data that reflect those disturbances and interruptions that are irrelevant. It is necessary to pre-treat the data files to exclude these “damaged” segments. The resulted “data-mining” algorithm is generic and can be used to deal with any large and complex data sets.

The first step of this research is to apply Ensemble Empirical Mode Decomposition (EEMD) and generate Intrinsic Mode Functions (IMFs) using data set available. Using IMFs frequency signals and standard mathematical concepts (Hilbert Transform), the anomalies are detected, and this gives the intrinsic relationship between the features used (equipment, weather, thermal resistance). Second, the frequency signals from iterated IMFs are constructed back into the original database using inverse Hilbert Transform (IHT), and the constructive data is utilized using semi-supervised machine learning algorithms to automate the process of pattern detection, cluster analysis, energy loss remediation and peak analysis.

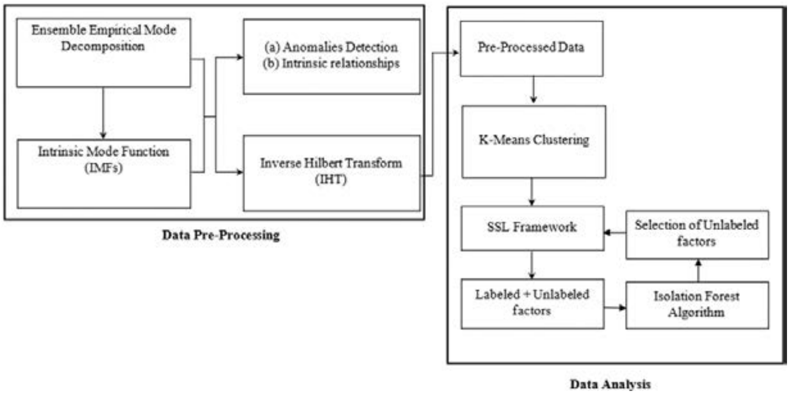
Ensemble Empirical Mode Decomposition
Intrinsic Mode Function (IMFs)
(a) Anomalies Detection
(b) Intrinsic relationships
Inverse Hilbert Transform (IHT)
Data Pre-Processing
Pre-Processed Data
K-Means Clustering
SSL Framework
Selection of Unlabeled factors
Labeled + Unlabeled factors
Isolation Forest Algorithm
Data Analysis

Figure 1 Proposed EEMD-SSL Framework

4. Preliminary Results and Discussion

Ideally, for a given data file, the EMD method returns a set of IMFs in separate frequency ranges. Practically, since each data file may be too large to be processed computationally, we need to divide the data into small segments so that each can be computed efficiently. To deal with the boundary effects properly, for each data segment, we include, from neighboring segments, an extra but much smaller subset of data points at both ends of the segment, forming the corresponding boundary sets. After performing the EMD calculations, only the IMFs within the original data segment are kept, while those associated with the boundary sets are disregarded. For a given data segment, the resulting IMFs usually depend on the choices of the sizes of the segment and the boundary sets. In particular, the larger the boundary sets, the more accurate the IMFs, but the amount of the computation will also increase. Our proposed procedure for analyzing large data sets thus consist of performing the EEMD to obtain different IMFs, calculating the amplitudes and frequencies of the IMFs that are deemed to reveal the dynamical evolution of the underlying system, and performing suitable statistical analyzes. Preliminary Results and Discussion

The research team completed a preliminary analysis on both EMD and EEMD to examine the differences and to modify the algorithm for the dataset collected. The data set is collected from the Energy Information System (EIS) and contains data at different time intervals (1 min, 15 min, hourly, daily and yearly). To perform the preliminary analysis on EEMD, the research team selected the medium frequency data that involves daily totals electricity consumption data for the whole year for commercial building. Figures 2 below depict the result of the EEMD and EMD for the same dataset for a year (365 data points) on energy consumption, and it is clear that EEMD has better results than EMD in handling oscillations. With x-axis on the number of days in a year (365), Figures 2 show the IMFs (1-5) with raw signals at the top. For instance, on IMF-2, the process of iteration looks different in both of the graphs where EEMD tends to display more intrinsic signals. Processing further, IMFs 3 and 4 show how EMD generates high oscillations when compared with the similar frequency of EEMD.

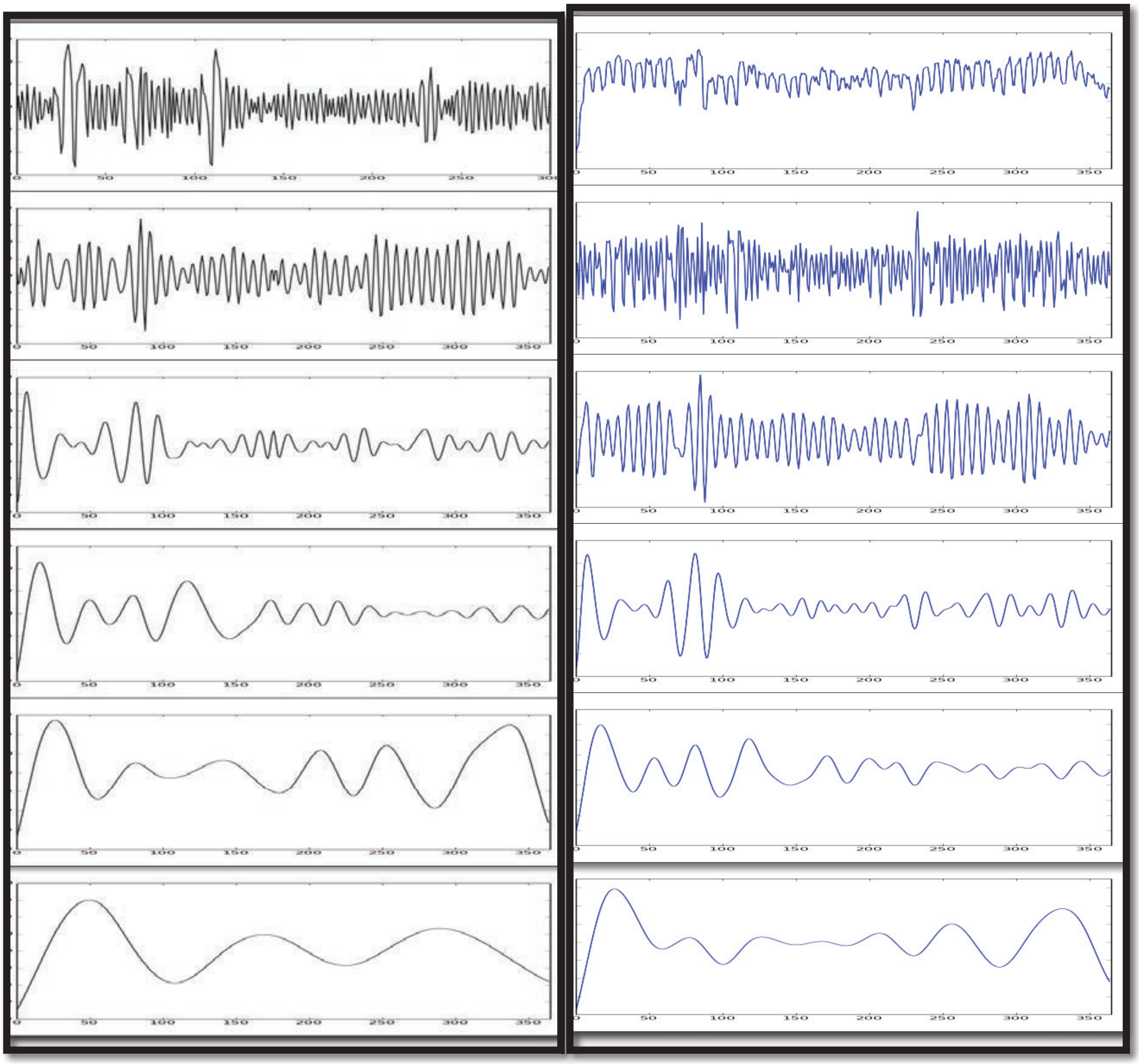

Figure 2 IMFs of EMD and EEMD for energy consumption database

Note: X axis is the number of days in a year (365) and Y-axis is the IMF signals (from 1 to 5)

The final IMF shows a clearer view of how EEMD prevents the oscillation variations and flattens most of the lower value or zero data points. Thus, EEMD overcomes the difficulties of the “flat” parts that the EMD algorithm deviates to higher oscillations. Also, EEMD helps unlabeled or unstructured data set that can give us unknown pattern, behavioral changes and intrinsic relationships between devices [24]. The anomaly detection will be performed using the process above (EEG data). Thus, EEMD is utilized to develop a better comprehensive framework to detect anomalies that stand to the energy loss.

5. Proposed EEMD-SSL Framework

Hilbert transform is a special case of harmonic analysis that undergoes convolution on the data u(t) [where u is the data which is time dependent] to produce discrete data in the frequency domain to undergo further case study in EEMD. However, this domain displaced data can be retraced using Inverse Hilbert Transform (IHT) [25]. It is due to the special case of a harmonic conjugate of this kernel function H (inverse), (h (u)) = -h [24]. Thus, getting back the original data necessary to implement paralleled study of Semi-supervised learning framework (SSL). It is a real-time energy demand and supply framework that would accurately estimate the energy consumption of building clusters by predicting the energy demand and supply for every cluster through the extensive implementation of semi-supervised learning techniques [25] (see following sections for existing work of PI on SSL). Through the learning process, the machine could predict energy loss percentage more accurately by analyzing unlabeled factors that account for energy losses. The unlabeled factors or features are selected using isolation technique that can select the confident unlabeled factors that improve the accuracy of the framework. With the proposed EEMD-SSL framework, dynamic model can be developed that would be an important decision-making and marketing strategy tool for the energy suppliers.

5.1 Feature selection of Unlabeled factors using Isolation Forest Algorithm

The proposed EEMD-SSL framework has two stages of anomalies. The PI proposed EEMD based anomaly method to identify the anomalies as a first step in analyzing the data. During semi-supervised learning framework, the research involves both labeled and unlabeled factors to elevate the accuracy of clustering by providing training set with no anomalies. The second stage of anomaly detection is during the selection of unlabeled factors. [26] surveys 3 different types of anomaly detection such as unsupervised clustering, supervised classification, and semi-supervised recognition and identified semi-supervised detection as the most effective method with greater accuracy. Ensemble based minimum margin active learning is a simple novel method for detecting anomalies using unsupervised learning [26], [27]. To detect anomalies and to select high confident unlabeled factors, a new and novel isolation forest algorithm is adopted that is faster and has greater accuracy than ORCA and Random forest [28]. The most important assumptions of Isolation forest algorithm is that the anomalies are a minority, and the attribute values are different from each other[28]. Isolation forest algorithm is best suited in high dimensionality [28] where the presence of irrelevant attributes (unlabeled data) is high (the case in big data) and in a situation where training set requires no anomalies, which is an important requirement for SSL.

The unlabeled factors integrated with SSL algorithm is selected by using Isolation forest technique in this research. After training the machine with labeled factors, the confident unlabeled factors from the isolation forest results are integrated with SSL framework as the labeled data. Like SSL, Isolation forest has three stages that include training, testing, and evaluation. The method builds an iTrees for the consumption data set, and then normal consumption patterns are clustered at the top end of the tree whereas the anomalies stay at the roots. The advantage of iTrees is that it can provide results of high dimensionality and efficiency with small sub-sampling data.

6. Conclusion

The EEMD-based algorithms that the team plans to develop have the potential to optimize energy prediction/forecast and align energy design and operation. The algorithms will create a platform that leads to a fully automated method to detect dynamical anomalies from large and complex data sets. It is anticipated that the proposed method would detect a large number of anomalies from generic large and complex data sets, which are not detectable using traditional methods. It also provides a superior test ground for probing into the emergence and evolution of anomalies through detailed analysis using methods from nonlinear dynamics, statistics, and statistical physics. Special features associated with different types of dynamical activities will be identified, with the goal to exploit the predictive power of anomalous behaviors. The detection will lead to the development of energy control systems that could be used to optimize energy design and operation. With the optimized EEMD-SSL based method, anomalies can be detected reliably for all the channels. The issue of the spatiotemporal evolution of these dynamical events can then be addressed.

The framework will be extending to a complex infrastructure system, and the correlation patterns of the distinct anomalous events can be identified.

References

[1] N. E. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N.-C. Yen, C. Tung, and H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proc. R. Soc. London. Ser. A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[2] T. Yalçınkaya and Y.-C. Lai, “Phase Characterization of Chaos,” *Phys. Rev. Lett.*, vol. 79, no. 20, pp. 3885–3888, 1997.

[3] S. B. Environments, “Post-occupancy evaluation of energy and indoor environment quality in green buildings : a review,” *3rd Int. Conf. Smart Sustain. Built Environ.*, pp. 1–7, 2009.

[4] D. Majcen, L. C. M. Itard, and H. Visscher, “Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications,” *Energy Policy*, vol. 54, pp. 125–136, 2013.

[5] A. Ahmed, N. E. Korres, J. Ploennigs, H. Elhadi, and K. Menzel, “Mining building performance data for energy-efficient operation,” *Adv. Eng. Informatics*, vol. 25, no. 2, pp. 341–354, 2011.

[6] S. Petersen and S. Svendsen, “Method for simulating predictive control of building systems operation in the early stages of building design,” *Appl. Energy*, vol. 88, no. 12, pp. 4597–4606, 2011.

[7] P. Raftery, M. Keane, and A. Costa, “Calibrating whole building energy models: Detailed case study using hourly measured data,” *Energy Build.*, vol. 43, no. 12, pp. 3666–3679, 2011.

[8] Z. O’Neill, B. Eisenhower, S. Yuan, T. Bailey, S. Narayanan, and V. Fonoberov, “Modeling and calibration of energy models for a DoD building,” *ASHRAE Trans.*, vol. 117, no. 860, pp. 358–365, 2011.

[9] C. Ghiaus, “Experimental estimation of building energy performance by robust regression,” *Energy Build.*, vol. 38, no. 6, pp. 582–587, 2006.

[10] N. Djuric and V. Novakovic, “Identifying important variables of energy use in low energy office building by using multivariate analysis,” *Energy Build.*, vol. 45, pp. 91–98, 2012.

[11] M. Kavgic, A. Mavrogianni, D. Mumovic, a. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic, “A review of bottom-up building stock models for energy consumption in the residential sector,” *Build. Environ.*, vol. 45, no. 7, pp. 1683–1697, 2010.

[12] C. C. Koopmans and D. W. te Velde, “Bridging the energy efficiency gap: using bottom-up information in a top-down energy demand model,” *Energy Econ.*, vol. 23, no. 1, pp. 57–75, 2001.

[13] P. De Wilde, “The gap between predicted and measured energy performance of buildings: A framework for investigation,” *Autom. Constr.*, vol. 41, pp. 40–49, 2014.

[14] G. Dall’O', L. Sarto, N. Sanna, and A. Martucci, “Comparison between predicted and actual energy performance for summer cooling in high-performance residential buildings in the Lombardy region (Italy),” *Energy Build.*, vol. 54, pp. 234–242, 2012.

[15] P. Raftery, M. Keane, and J. O’Donnell, “Calibrating whole building energy models: An evidence-based methodology,” *Energy Build.*, vol. 43, no. 9, pp. 2356–2364, 2011.

[16] L. Wang, P. Mathew, and X. Pang, “Uncertainties in energy consumption introduced by building operations and weather for a medium-size office building,” *Energy Build.*, vol. 53, pp. 152–158, 2012.

[17] A. Dodoo, L. Gustavsson, and R. Sathre, “Building energy-efficiency standards in a life cycle primary energy perspective,” *Energy Build.*, vol. 43, no. 7, pp. 1589–1597, 2011.

[18] M. Ryghaug and K. H. Sørensen, “How energy efficiency fails in the building industry,” *Energy Policy*, vol. 37, no. 3, pp. 984–991, 2009.

[19] L. Pérez-Lombard, J. Ortiz, and C. Pout, “A review on buildings energy consumption information,” *Energy Build.*, vol. 40, no. 3, pp. 394–398, 2008.

[20] N. E. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N.-C. Yen, C. Tung, and H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proc. R. Soc. London. Ser. A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[21] Y. Lai, “Analytic signals and the transition to chaos in deterministic flows,” *Phys. Rev. E*, vol. 58, no. 6, pp.

6911–6914, 1998.

[22] and J. D. Z. C. H. Liang, Q. H. Lin, "Application of the empirical mode decomposition to the analysis of esophageal manometric data in gastroesophageal reflux disease," no. 302, pp. 620–623, 2000.

[23] J. C. Nunes, Y. Bouaoune, E. Delechelle, O. Niang, and P. Bunel, "Image analysis by bidimensional empirical mode decomposition," *Image Vis. Comput.*, vol. 21, no. 12, pp. 1019–1026, 2003.

[24] J.-P. Tang, D.-J. Chiou, C.-W. Chen, W.-L. Chiang, W.-K. Hsu, C.-Y. Chen, and T.-Y. Liu, "A case study of damage detection in benchmark buildings using a Hilbert-Huang Transform-based method," *J. Vib. Control*, vol. 17, no. 4, pp. 623–636, 2010.

[25] H. Naganathan, W. K. Chong, and X. Chen, "Semi-supervised Energy Modeling (SSEM) for Building Clusters Using Machine Learning Techniques," *Procedia Eng.*, vol. 118, pp. 1189–1194, 2015.

[26] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2. pp. 85–126, 2004.

[27] K. Yamanishi, J. I. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Min. Knowl. Discov.*, vol. 8, no. 3, pp. 275–300, 2004.

[28] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2008, pp. 413–422.