Sujet de thèse

Co-clustering de séries temporelles pour l'aide à l'évaluation des performances énergétiques de bâtiments

Mots clés : *énergie, bâtiment, co-clustering dynamique, séries temporelles de grande dimension*
Lieu de travail : *Université Gustave Eiffel, laboratoire GRETTIA, Champs-sur-Marne*
Profil : *Master 2 ou Ingénieur dans les champs disciplinaires liés à la science des données (statistique, informatique)*
Pré-candidature : *envoyer cv et lettre de motivation à Allou Samé :allou-badara.same@univ-eiffel.fr*

Contexte

Le secteur du bâtiment, reconnu pour avoir la plus grande consommation d'énergie, constitue le levier d'action principal des politiques actuelles d'économie d'énergie. Dans ce cadre, la mise en place d'outils de comparaison de panels de bâtiments en termes de performances énergétiques est utile pour un meilleur ciblage des politiques de rénovation. Les méthodes généralement utilisées visent à construire des indicateurs de consommation annuelle d'énergie normalisés par rapport à la surface des bâtiments, aux caractéristiques statiques, à la température ou à d'autres facteurs tels que les périodes d'opération [1]. D'autres stratégies sont fondées simplement sur un étiquetage manuel plus ou moins détaillé des bâtiments selon l'usage qui leur est réservé (ex. résidentiel, école, bureau, commerce) [2]. Ces méthodes ne tiennent cependant pas compte de la dynamique d'utilisation ou d'opération des bâtiments, ce qui peut entrainer un biais dans l'analyse des performances énergétiques.

Objectifs et problématique

Cette proposition de thèse s'inscrit dans le cadre de l'exploitation de données temporelles massives issues du fonctionnement réel de bâtiments pour analyser de manière fine leurs performances énergétiques. Les données en question sont formalisées par des séries temporelles numériques à dimensionnalité élevée collectées sur chaque bâtiment via des smart meters (ex. série horaire de consommation d'électricité ou de gaz), ainsi que de données décrivant l'environnement ambiant de bâtiments (ex. température, humidité) collectées via des équipements dédiés. La thèse vise à proposer des méthodes avancées de classification non supervisée de ces longues séries temporelles pour le regroupement de bâtiments en termes de performances énergétiques, et qui pourront aussi fournir un espace de représentation simplifié des données. La spécificité de ce problème réside dans la dimensionnalité élevée des séries considérées ainsi que dans la diversité de leurs variations, liées à de multiples facteurs observables ou non observables, qu'il faudra prendre en compte dans la construction d'espaces de représentation cohérents avec l'objectif de caractérisation énergétique. Ce caractère complexe des données rend les méthodes classiques de partitionnement de séries temporelles, notamment celles à base de distances (euclidienne, dynamic time warping, corrélation) [2] inadaptées. Des études récentes [3, 4] ont conduit à proposer des stratégies en deux étapes : simplification de chaque série temporelle à l'aide d'un nombre réduit de profils journaliers, puis partitionnement des bâtiments en utilisant les données codées via le dictionnaire formé par les profils-types. La conception d'un modèle probabiliste unifié regroupant ces deux étapes peut contribuer à mieux optimiser le regroupement de bâtiments en classes homogènes du point de vue des performances énergétiques.

Méthodologie

Pour aborder ce problème de classification et de réduction de la dimensionnalité d'un panel de séries temporelles, cette thèse se focalisera sur les méthodes de classification croisée (co-clustering). Si les méthodes de classification classiques visent à partitionner les lignes d'un tableau de données, les méthodes de co-clustering visent quant-à-elles à partitionner simultanément ses lignes et ses colonnes ; ce qui revient aussi à organiser les données en blocs homogènes [5]. Dans le cadre de cette thèse, les lignes pourront être associées à des séries temporelles symbolisant les bâtiments et les colonnes à des périodes temporelles judicieusement choisies. La thèse vise à étendre les travaux récents dans le

domaine de la classification croisée [6], à des modèles tenant compte explicitement du caractère temporel des données. Le modèle à blocs latents (statiques) qui est une référence en la matière [5] pourra ainsi être étendu à un **modèle à blocs latents dynamiques pour pouvoir gérer la dépendance temporelle des données entre les blocs et au sein des blocs**. Une étude comparative des performances de cette approche par rapport aux méthodes effectuant simultanément et de manière non supervisée la classification et l'apprentissage de représentation à l'aide du deep learning et des auto-encodeurs [7] sera également menée. Il s'agira aussi d'évaluer la pertinence des clusters de bâtiments issus de ces méthodes, au regard des performances énergétiques de ces derniers.

La mise en application des méthodes développées au cours de cette thèse sera réalisée à partir de sources de données réelles qui sont rendues disponibles dans le cadre académique. On peut faire référence par exemple aux bases de données « Build Smart DC », « CER smart meter data » ou encore « EnerNOC Green Button Data » qui répertorient la consommation d'énergie de multiples bâtiments ainsi que les caractéristiques statiques de ces derniers. L'une des tâches de la thèse sera ainsi consacrée à la recherche, la compilation et la mise en forme de telles bases de données.

Les travaux abordés dans cette thèse, qui portent sur l'aide à l'évaluation des performances énergétiques de bâtiments, viennent en complément de recherches récentes menées au laboratoire Grettia sur la mise au point de méthodes non supervisées d'apprentissage du comportement d'occupants de bâtiments à partir de données massives [8, 4].

Programme prévisionnel

L'organisation prévue pour cette thèse est la suivante :

- Phase 1 : étude bibliographique sur les méthodes de benchmarking des performances énergétiques de bâtiments à partir de données massives et constitution de bases de données décrivant la dynamique énergétique de bâtiments ;
- Phase 2 : mise au point de méthodes de co-clustering à blocs dynamiques pour la classification et la représentation d'un ensemble de séries temporelles associées à des bâtiments ;
- Phase 3 : application des nouvelles méthodes proposées à des données réelles, et comparaison de celles-ci aux méthodes issues de l'état de l'art récent (deep learning).

Quelques références

[1] W. Chung, Y. Hui, and Y. M. Lam (2006). Benchmarking the energy efficiency of commercial buildings, Applied energy 83 (1) (2006) 1–14.

[2] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah (2015). Time-series clustering; a decade review. Information Systems 53, pp. 16-38.

[3] J. Y. Park, X. Yang, C. Miller, P. Arjunan, and Z. Nagy (2019). Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset, Applied Energy 236, pp. 1280–1295.

[4] M. Leyli-Abadi, A. Samé, L. Oukhellou, N. Cheifetz, P. Mandel, C. Féliers, and O. Chesneau (2019). Mixture of Joint Nonhomogeneous Markov Chains to Cluster and Model Water Consumption Behavior Sequences. ACM Transactions on Intelligent Systems Technologies 10, 6, 71.

[5] G. Govaert and M. Nadif M. (2014). Co-clustering. Computer Engineering. Wiley, New York.

[6] C. Bouveyron, L. Bozzi, J. Jacques, and F.-X. Jollois (2018). The functional latent block model for the co-clustering of electricity consumption curves. Journal of the Royal Statistical Society Series C, Royal Statistical Society, vol. 67(4), pp. 897-915.

[7] Q. Ma, C. Chen, S. Li, and G. W. Cottrell (2021). Learning Representations for Incomplete Time Series Clustering. Proceedings of the AAAI Conference on Artificial Intelligence, 35(10), pp. 8837-8846.

[8] L. Bonfils, A. Samé, and L. Oukhellou (2021). Dynamic clustering and modeling of temporal data subject to common regressive effects. Proceedings of the 29th European Symposium on Artificial Neural Networks (ESANN).