BILINGUAL PUBLISHING GROUP

Journal of Computer Science Research

https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Learning Dominant Urban Flows around High-Rise Buildings with Data-Driven Balance Models

Zhiyu Huo

Department of Civil Engineering, IMPERIAL COLLEGE LONDON, London, SW7 2AZ, UK

ABSTRACT

This thesis develops a data-driven dominant balance model to recognise and cluster the flow pattern blowing through a high-rise building in an urban area under neutral atmospheric conditions. To be consistent with the governing equation used in simulations, the Reynolds-Averaged Navier-Stokes (RANS) equation is selected as the governing equation. It is divided into six sub-parts based on the physical meanings of each term in RANS. The time-averaged simulation results are used as the data set basis for further machine learning and clustering. The approach used to achieve the final dominant balance models consists of knowledge from fluid mechanics, statistics and programming. Knowledge from fluid mechanics is mainly used for proposing governing equations and interpreting the final outcomes, whereas the knowledge from programming is used for script writing and program running. Finally, the knowledge from statistics is the key for algorithms to achieve the clustering and dominant balance model acquirement. This approach includes the finite difference method, Gaussian mixture models (GMM), singular value decomposition and sparse principal component analysis (SPCA). The finite difference method is used for approximating the derivatives in RANS, which works as a post-processing step. GMM are trained by using randomly subsampled points and applied for the clustering of the processed data points. A drawback of yielding overlapping and trivial clusters of GMM is spotted and SPCA is applied as the solution to trivial results, using regularisation to proceed with a sparse approximation for excessive cluster elimination. The final data-driven dominant balance models are obtained and visualised by generating two tables for two cases.

Keywords: Machine learning; Urban flows; Fluid mechanics

*CORRESPONDING AUTHOR:
Zhiyu Huo, Department of Civil Engineering, IMPERIAL COLLEGE LONDON, London, SW7 2AZ, UK; Email: huozhiyuhh@163.com

ARTICLE INFO
Received: 31 July 2024 | Revised: 8 August 2024 | Accepted: 10 August 2024 | Published Online: 20 August 2024
DOI: https://doi.org/10.30564/jcsr.v6i4.6984

CITATION
Huo, Z., 2024. Learning Dominant Urban Flows around High-Rise Buildings with Data-Driven Balance Models. Journal of Computer Science Research. 6(4): 1–18. DOI: https://doi.org/10.30564/jcsr.v6i4.6984

COPYRIGHT
Copyright © 2024 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (https://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

1.1 Overview

Nowadays more than half of the population lives in urban areas and this number is expected to continuously increase in the future [1]. This makes the development of cities important because urban areas are the centre of energy consumption and heat and pollution emissions. An increasing number of articles and technical reports are focusing on urban areas, aiming for a better understanding of the urban climate and an improvement in the living standards of urban residents.

Many aspects of urban areas have been focused on by academics, such as pollution and waste dispersion and the urban heat island effects. To mitigate these problems, it is imperative to understand the flow patterns in complicated urban areas, so that making the buildings contribute to the heat and pollution transportation processes. Many experimental and computational approaches have been applied to simulate the flow interacting with building geometries.

As a result of the continuously increasing population in urban areas, skyscrapers and other kinds of high-rise buildings have been constructed to house more people. From the logarithmic profile of the atmospheric boundary layer, it is obvious that the high-rise buildings will suffer from a faster air flow. Consequently, the interactions between these buildings and air flows are of great importance.

The experimental, numerical and other traditional methods for urban flow study are extremely demanding. It would normally take days to run a single simulation on sophisticated hardware such as a supercomputer. Moreover, the simulation results would not be able to illustrate the flow regimes and therefore, the understanding based on the simulation results would be limited – especially when buildings in urban areas have different heights and geometries.

However, the data-driven dominant balance model would be able to solve this problem of limited understanding. By using machine learning techniques, this model can identify different dominant terms of the Reynolds Averaged Navier Stokes Equation (RANS) in the urban flow field and cluster the regions into different clusters for further study and analysis.

To the best of my knowledge, no one has yet applied this method to urban flow analysis. Therefore, this research project will be the application and testing of a brand-new approach to this field of study. The urban flow model will be demonstrated in the following background section, which is the work of one of the last-year graduates. The methodology section focuses on the methods and principles covered in the data-driven dominant balance model. Finally, the results and discussion section pays attention to the results of the model. All the computations and model training processes are completed in MATLAB and Python. Attaching the scripts to this paper would be unnecessary, with further reasoning provided in the appendix.

1.2 Background

The urban model used in this research is the staggered cubic arrays and high-rise buildings that is generated by colleagues. Two cases with different atmospheric conditions have been set and simulated in Zhang's paper. The basis of this research project is the simulation results of the case with neutral atmospheric conditions.

Case in neutral atmospheric conditions

The domain size of the urban geometry configuration is 3200 m × 640 m × 480 m for x, y, and z directions, respectively and it is set in the Cartesian coordinate system. The layout of the geometry is shown in **Figure 1**. From the upper elevation view, it is obvious that a high-rise building is located in the front of the domain. It has the dimensions 40m × 40m × 240m in length, width and height, respectively. Putting the high building in the front is useful for forming a clear wake development in the rest of the domain. All other buildings have the same dimensions but only 80m in height. The lower graph in **Figure 1** shows the plane view of the urban model, which is staggered cubic arrays. The streets between buildings are modelled as canyons extending the entire length of the y-direction (640 m). Furthermore, there

is a relatively large area at the end of the domain to allow the air to smoothly flow out. The final resolution of the geometry configuration is Nx × Ny × Nz = 640 × 128 × 192.

The simulation dataset for this research is based on this geometry and under neutral atmospheric conditions. 'Neutral atmospheric conditions' refers to the environmental lapse rate equalling the dry adiabatic rate in dry air. For example, if a parcel of air is lifted through a neutral layer, the temperature and pressure of the parcel will be identical to the temperature and pressure of the surrounding air at every height and will always be in equilibrium with the environment [2].

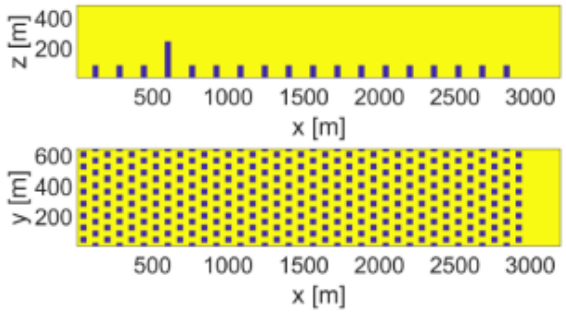

Figure 1. Schematic diagram of the urban model geometries. The upper graph shows the elevation view and the lower one shows the plane view.

Data-driven dominant balance models

Exploration of the data-driven dominant balance models on the urban fluid mechanics is inspired by an article written by Callaham et al. in 2021, which introduces the application of a data-driven approach to dominant balance analysis on a variety of physical processes, including the boundary layer in transition to turbulence, the nonlinear optical pulse propagation, geostrophic balance in the Gulf of Mexico and a generalised Hodgkin-Huxley model. Take the boundary layer in transition to turbulence for example, the general steps are introducing the governing equation, plotting the Reynolds-averaged fields, clustering and obtaining the dominant balance models. Finally, a general graph of the clustering regions is plotted for better visualisation and discussion. These steps are shown in **Figure 2**.

The data-driven approach to dominant balance analysis generalises traditional methods in several critical directions. Firstly, this approach does not depend on any explicit assumption of asymptotic scaling. Secondly, the clustering method provides pointwise estimates of the spatiotemporally local dominant balance not afforded by traditional scaling analysis in complex geometries. Moreover, it provides an objective, reproducible approach to testing these hypotheses while many dominant balance regimes have been proposed or assumed based on heuristic or intuitive arguments. Finally, the probabilistic Gaussian mixture modelling framework is fully compatible with the relative nature of dominant balance analysis, providing natural estimates of uncertainty in the identified balance [3].

1.3 Aims and objectives

The objectives of this study are to:

(a) Learn and understand the Gaussian Mixture Model and Sparse Principal Component analysis and achieve model training and clustering based on Zhang's high-rise building simulation data sets.

(b) Apply the finite difference method to discretise and approximate the terms in the Reynolds Averaged Navier Stokes equation.

(c) Train GMM and apply it for clustering the data sets. Then, use l_1 regularisation in SPCA to proceed with a sparse approximation to reconstruct the clustering results.

(d) Interpret and compare the difference between the results before and after sparse approximation and interpret the results. Finally, dominant balance models are obtained after sparse approximation.

2. Methodology

In this section, the methods and theories that have been covered in the research will be illustrated, including the governing equation determination of the urban flow field, the finite difference (FD) method, Gaussian Mixture Models, Sparse Principal Component Analysis as well as the Dominant Balance Models. The following sub-sections will not only introduce the principles and equations from each method or theory but also the related MATLAB or Python commands.

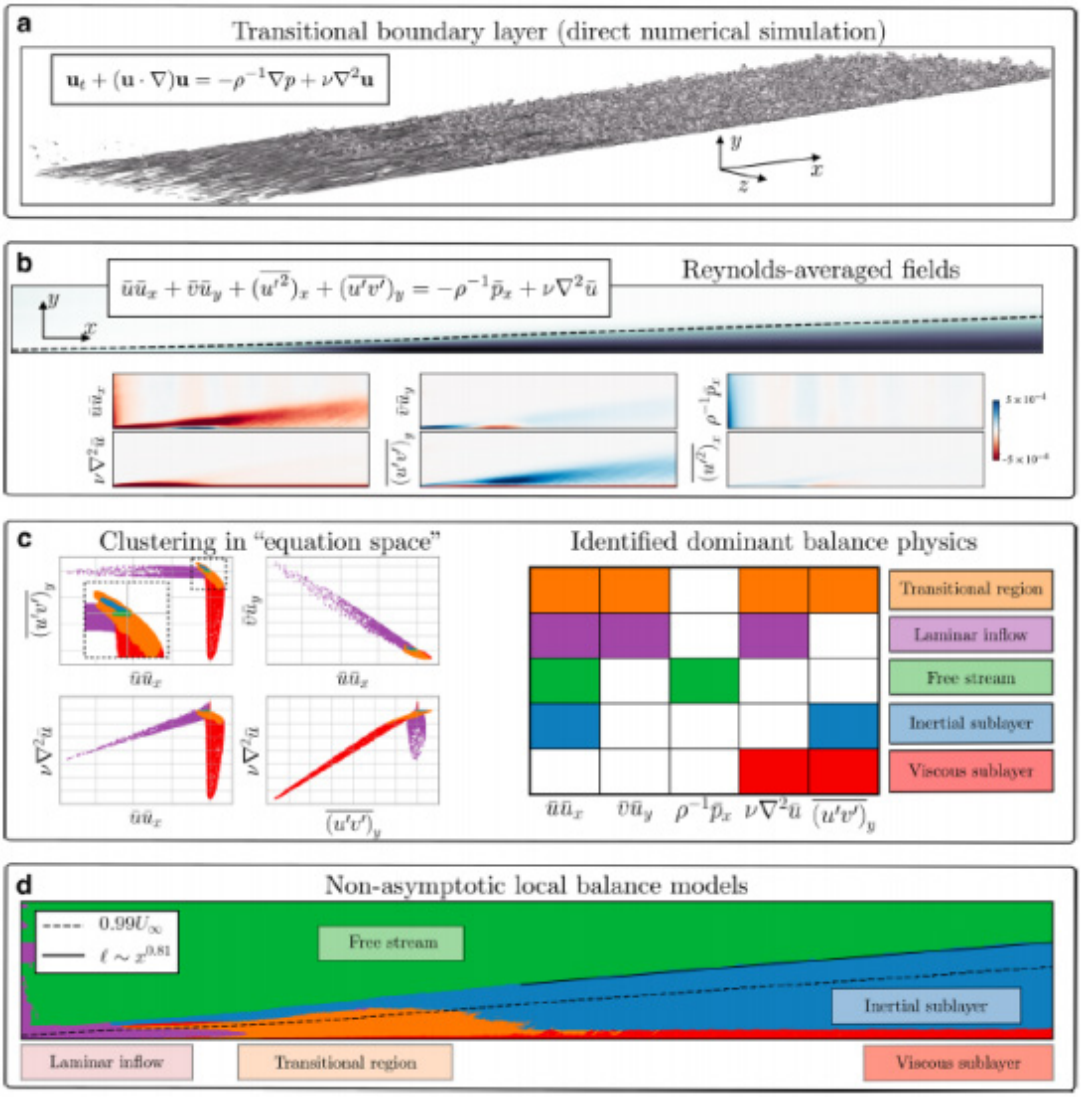

Figure 2. Schematic of the dominant balance identification procedure applied to a turbulent boundary layer.

Source: Brunton [3].

2.1 Governing equation

Navier-Stokes equations are certain partial differential equations that describe the motion of Newtonian fluid substances. They can be used for modelling weather, ocean currents and waves, pollution dispersions and air flows around an aerofoil. Mathematically speaking, the essence of the equations is the conservation of momentum and the conservation of mass, which are the governing equations of the simulation software. The equations in full can be written as:

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+w\frac{\partial u}{\partial z}=-\frac{\partial p}{\partial x}+\frac{1}{Re}\nabla^2 u \quad (1)$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+w\frac{\partial v}{\partial z}=-\frac{\partial p}{\partial x}+\frac{1}{Re}\nabla^2 v \quad (2)$$

$$\frac{\partial w}{\partial t}+u\frac{\partial w}{\partial x}+v\frac{\partial w}{\partial y}+w\frac{\partial w}{\partial z}=-\frac{\partial p}{\partial x}+\frac{1}{Re}\nabla^2 w \quad (3)$$

The equations (1), (2) and (3) are Navier-Stokes equations in the x, y and z-direction, respectively. The velocity u, v, and w are in four dimensions which are x, y, z and time. However, the time dimension is not useful because the primary objective is to identify different flow regimes of the high-rise building model, hence time-varying variables are not

necessary. Therefore, the governing equation can be simplified into the Reynolds-Averaged Navier-Stokes equation (RANS). The 3-dimensional RANS equation for horizontal velocity u can be written as:

$$\overline{u}\frac{\partial \overline{u}}{\partial x}+\overline{v}\frac{\partial \overline{u}}{\partial y}+\overline{w}\frac{\partial \overline{u}}{\partial z}=-\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x}+\nu\nabla^2\overline{u}-\frac{\partial \overline{u'u'}}{\partial x}-\frac{\partial \overline{u'v'}}{\partial y}-\frac{\partial \overline{u'w'}}{\partial z} \quad (4)$$

The overline of each variable represents the time-average property. The new term $\left(-\frac{\partial \overline{u'u'}}{\partial x}-\frac{\partial \overline{u'v'}}{\partial y}-\frac{\partial \overline{u'w'}}{\partial z}\right)$ is the turbulence term, representing the Reynolds stresses. The same decomposition method is applied to velocities and as well. The RANS equation will be the governing equation throughout the research.

The RANS equation has also been divided into six parts for further discussion and analysis.

$$\{\overline{u}\overline{u}_x+\overline{v}\overline{u}_y\}_{①}+\{\overline{w}\overline{u}_z\}_{②}=\{\rho^{-1}\overline{p}_x\}_{③}+\{\nu\nabla^2\overline{u}\}_{④}-\left\{\left(\overline{u'^2}\right)_x-\left(\overline{u'v'}\right)_y\right\}_{⑤}-\left\{\left(\overline{u'w'}\right)_z\right\}_{⑥} \quad (5)$$

Horizontal advection term, vertical advection term, pressure gradient, viscous term, horizontal Reynolds stresses and vertical Reynolds stresses are labelled as 1–6 respectively.

2.3 Gaussian mixture models

One of the popular unsupervised clustering methods is known as finite mixture models. These models are a mixture of Gaussian distributions with different means and covariances, so this method is named the Gaussian mixture model (GMM) [5]. A one-dimensional example from C. Bishop is used here to show why GMM is useful [4].

The red line in **Figure 3** has three dominant clumps, which is impossible to capture using a single Gaussian distribution. However, a superposition of three Gaussian distributions (blue lines) can significantly improve the characterisation of the probability function $p(x)$. Therefore, almost any continuous density can be approximated by using a sufficient number of Gaussians and adjusting their means and covariances [4]. The expression for a superposition of K Gaussian densities can be written in the following form:

$$p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k) \quad (6)$$

Each Gaussian distribution $N(x|\mu_k, \sum_k)$ is called a component of the mixture, having a specific mean μ_k and covariance $\sum_k$. The term π_k is the mixing coefficient [4].

In this research, the number of Gaussian distributions is set as 6, representing six different clusters of the mixture model. The learned covariances for each cluster can be interpreted in terms of active and inactive terms. Active terms will have a certain value whereas the corresponding inactive terms are near-zero, which is negligible. Data beyond the original inputs can efficiently be assigned to a balance model using the trained GMM [3].

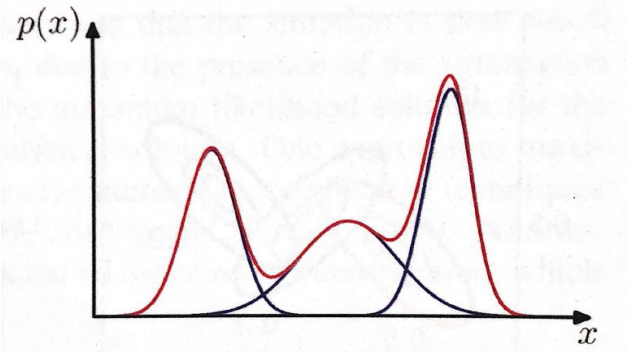

Figure 3: One-dimensional example of Gaussian mixture distribution.

Source: Callaham [4].

2.4 Singular value decomposition

Principal component analysis (PCA) is one of the most widely used techniques for taking high-dimensional data and trying to understand it with dominant patterns and correlations. As the basis of PCA, the matrix factorisation method Singular Value Decomposition (SVD) should be introduced in advance. SVD is one of the most popular approaches used in numerical linear algebra for data processing. It can be used in data reduction, dimensionality reduction and as the foundation of machine learning. It uses simple linear algebra to decompose the data set, ob-

taining interpretable and understandable features that can be used to build models.

A large data set $X \in C^{m \times n}$ can be written in a matrix form with rows and columns:

$$X = \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & \cdots & | \end{bmatrix} \tag{7}$$

Each long vector x represents elements of the large data set. For example, a large face database X with data representing a lot of people's facial information x. Each piece of information x is in n dimensions, such as hair, nose, eyes, ears and mouth. The SVD is a unique matrix decomposition that exists for every complex-valued, non-squared matrix X.

$$X = U\Sigma V^T \tag{8}$$

$U \in C^{n \times n}$ and $V \in C^{m \times m}$ are called unitary matrices with orthonormal columns. $\Sigma \in R^{n \times m}$ is a matrix with real, non-negative entries on the diagonal and zeros off the diagonal. Although the diagonal matrix $\sum$ has the same dimension as X, there are only m singular values in $\sum$, which means the elements below the m^{th} singular value are zero and the vectors in matrix U after m^{th} column would be trivial. Therefore, economy SVD is used, acting as a simplification of the full SVD.

$$X = \hat{U}\hat{\Sigma}V^T \tag{9}$$

Where $\hat{U}$ and $\hat{\Sigma}$ are matrices after simplification, and $\hat{\Sigma} \in C^{m \times m}$.

2.5 Sparse principal components analysis

Principal component analysis (PCA) is a central use of SVD, which provides a data-driven, hierarchical coordinate system to represent high-dimensional correlated data. PCA pre-process the data by mean subtraction and setting the variance to unity before performing the SVD [5]. The computation of principal components (PCs) is the key of PCA. PCs are uncorrelated to each other but maximally correlated to the measurements. The number of PCs depends on the number of variables or dimensions of the measurements. The PCs computation can be divided into six steps:

First, compute the row-wise mean $\overline{\mathrm{x}}$ which is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

and then create an average matrix which is:

$$\overline{X} = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \bar{x} \tag{11}$$

Secondly, subtract the average matrix $\overline{\mathrm{X}}$ from the original data matrix X , which results in the mean-subtracted matrix B.

$$\mathrm{B} = \mathrm{X} - \overline{\mathrm{X}} \tag{12}$$

The covariance matrix of the rows of B is given by:

$$C = \frac{1}{n-1} B^T B \tag{13}$$

Compute the eigenvalues and eigenvectors of the covariance matrix C:

$$\mathrm{v_i^T B^T B v_i} \tag{14}$$

$$\mathrm{CV} = \mathrm{VD} \tag{15}$$

Where D is the eigenvalues and V is eigenvectors. Finally, the PCs can be computed by:

$$\mathrm{T} = \mathrm{BV} \tag{16}$$

Here, the matrix containing eigenvectors V is identical to the matrix V obtained from the SVD. Therefore, an alternative for PCs computation can be derived by carrying out SVD on the mean-subtracted matrix B, which can be written in the form of:

$$\mathrm{B} = \mathrm{U\Sigma V^T} \tag{17}$$

$$T = BV = U\Sigma V^T V = U\Sigma \tag{18}$$

The entire research is based on programming

in MATLAB and Python, hence the SVD and PCA computations should be completed by a series of commands. The SVD and PCA can be completed by using the commands '$[U, S, V] = svd(X)$' and '$[V, score, s2] = pca(X)$' in MATLAB, where S is the diagonal matrix $\sum$, the vector $s2$ contains eigenvalues of the covariance of x, the variable score contains the coordinates of each row of the mean-subtracted matrix B in the principal component directions [5]. Although the computation is done by simply using a single command, it is necessary and important to understand the computation processes of SVD and PCA otherwise the outcomes would be academically meaningless.

However, PCA suffers from the fact that each principal component is a linear combination of all the original variables, making it fragile with respect to outliers in measurements, hence the results are often difficult to interpret. A new method using the lasso to produce modified principal components with sparse loadings is introduced, which is called sparse principal component analysis (SPCA). PCA can be written in the form of a regression-type optimisation problem and SPCA is built on this with a quadratic penalty. The goal of using SPCA is to carry out a sparse approximation of the leading principal component.

For each i, denoted by $Z_i = U_i D_{ii}$ the i^{th} principal component and the coefficient λ is positive. The ridge estimates $\beta Tidge$ is given by the following:

$$\widehat{\beta}_{ridge} = \arg\min \left||Z_i - \boldsymbol{X}\beta|\right|^2 + \lambda\left||\beta|\right|^2 \tag{19}$$

However, the ridge estimation is not a general case to handle all kinds of data as parts of the data set have corrupted measurements that will spoil the estimation. Therefore, it is extended to a more general case by adding the L1 penalty to the equation (19).

$$\widehat{\beta} = \arg\min \left||Z_i - \boldsymbol{X}\beta|\right|^2 + \lambda\left||\beta|\right|^2 + \lambda_1\left||\beta|\right|_1 \tag{20}$$

Where $\lambda_1\left||\beta|\right|_1 = \sum_{j=1}^{p}|\beta_i|$ is the l1 -norm of β . Approximation to Vi can be expressed as = $||\hat{\beta}||$. Approximation to ith principal component is Xvi. A large enough value of λ_1 yields a sparse β and then a sparse vi. Thus, given a fixed λ is sufficient for solving all values of λ_1 by using the lasso-elastic net (LARS-EN) algorithm from Zou and Hastie [6]. The sparse approximation to any PC can be flexibly achieved [7].

Once a dominant balance regime is identified in a cluster, it has to be well-described by its direction of maximum variance. Additionally, the leading principal component of a cluster should have many non-zero entries and near-zero entries. Therefore, the SPCA is applied to the set of points in GMM clusters, taking the active terms and neglecting the inactive terms which represent the non-zero and near-zero entries, respectively. This process is a sparse approximation to the leading principal component.

2.6 Dominant balance models

The dominant balance model consists of dominant terms in the governing equation which balances each other. Dominant terms are identified by active and inactive terms (non-zero and near-zero entries) in the SPCA vector in the corresponding cluster. After applying l1 regularisation, each GMM cluster has a sparse approximation to its leading principal component. Different clusters may have the same sparsity pattern, which is considered to be part of the same dominant balance regime [3]. Points from all clusters with the same sparsity patterns are combined into the same dominant balance model. Once the dominant balance models are determined, the original domain can be divided into parts depending on the dominant physical processes in each local region.

3. Results and discussion

In this section, the results based on the methodology introduced in the previous section will be illustrated. Figures, interpretations, as well as the corresponding discussions based on the computation results, are covered. Both the plane view and side view of the urban area are included as this simulation is done in a 3-D domain. This section is divided into 4 subsections: the simulation results of the neutral atmosphere case, the clustering by Gaussian mixture models, SPCA reduction (sparse approximation) and the final dominant balance models.

3.1 Neutral atmosphere case

Before heading to the clustering and machine learning section, it is important to plot the outcomes from the simulation to examine the data set. If the data set is not reasonable, the final dominant balance model would be meaningless.

The simulation result is based on the governing equation of the Reynolds-averaged Navier-Stokes equation, hence the variables in the data set are all time-averaged and in 3-D. Due to the simulation settings, all variables apart from the mean velocities are defined at the centre of the cells in a mesh grid. Hence the velocities should be centred before plotting any graphs based on them. **Figure 4** consists of the graphs for six sections of the RANS equation. The graphs are plotted as a side view of the urban area at the domain width of 322.5m.

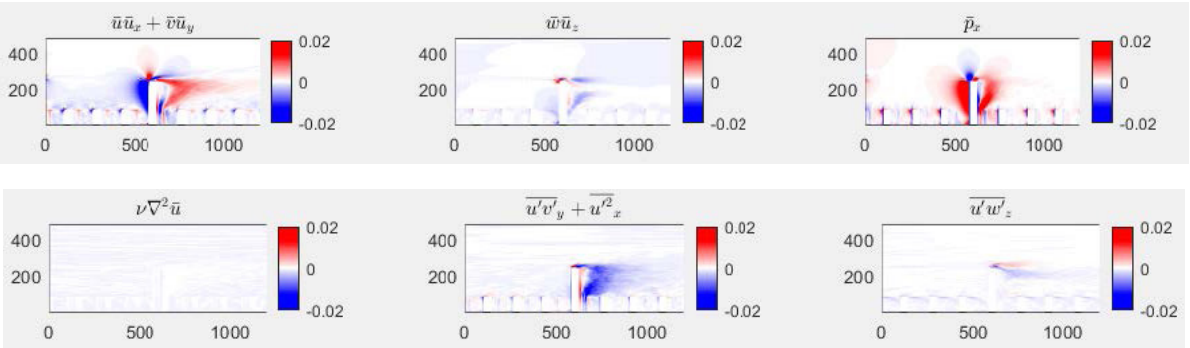

Figure 4. RANS terms at the side view of the urban area.

The entire urban area is long, but the model is segmented to 0 – 1200m because the building that we are interested in is the skyscraper. All small buildings are identical to each other. The colour bar is fixed in a range of ±0.02 to observe the magnitude difference of each term.

The horizontal advection terms have negative values in the front of the high-rise building because the building is impermeable and the flow will reflect when it reaches the building surface. In the wake region, the magnitude becomes positive but not large. This is because there has been a flow separation at the front edge of the tall building, the flow separation at the back edge of the tall building will not be as significant as the front edge. It is necessary to mention that there is a high velocity point at the top of the building near the edge. This is not only the result of the incoming horizontal flow but also the high-speed flow at the surface of the building after it reaches the stagnation point. A high-speed flow separation is then formed, so the horizontal advection terms at that point have a relatively large positive value. The vertical advection term shows a similar pattern, where the value at the top of the high building is large. Its magnitude in the wake regions is relatively larger than in other regions due to the turbulence.

The pressure gradient term shows a high-pressure region on the building front where the stagnation point is caused by horizontal incoming flow. The flow separation area at the top edge of the skyscraper shows a negative value of the pressure gradient, with respect to x. In the wake region at the back of the building, it is obvious that the pressure gradient is positive but much smaller than the stagnation point.

The viscosity term is near zero in the domain segment because the viscosity dominant area is small and near the surface. In comparison with the graph plotted, the region is too small to be shown in the graph. In other words, the grid Nx × Ny ×Nz = 640 × 128× 192 is too coarse to show the viscosity dominant region in this large-scale simulation. The horizontal Reynolds stresses have negative values in the wake region at the back of the high building and in the front of the first short building after the high building because the flow in the wake region is highly turbulent. The reason for the negative value is that the graph plotted is for (u'u' + u'v') whilst there is

an additional negative sign in front of the Reynolds stress terms in the RANS equation.

The vertical Reynolds stress term u'w' is having a similar pattern as horizontal Reynolds stress terms but its magnitude is nearer to zero. It is unusual that the vertical Reynolds stress has a much smaller value than the horizontal Reynolds stresses. A double-check of the simulation settings may be required but the details in running the simulation are beyond the scope of this paper.

The visualisation process also contains the plane view of the urban area. In **Figure 5**, the height of the domain is chosen as 98.75m, so that only the tall building itself is included in the slices.

Similar to **Figure 4**, six parts of the RANS equation are plotted in the plane view. The colour bars are fixed in the range of ±0.02 as well. For all graphs in **Figure 5,** the backgrounds are dot-like because of the existence of small buildings at the lower level. Flow separations are clearly demonstrated in these graphs. The general trend of magnitude of each term is the same as the side view. The viscous term and vertical Reynolds stress term are near zero. Another issue is observed by plotting the graphs in the plane view. Technically, the pressure gradient is supposed to be the same from the initial point to the end of the domain because the plane view is a slice of the z-axis.

However, the x graph in **Figure 5** indicates an increasing trend of the pressure gradient along the x-domain, which is against expectations. This problem may be caused by the initial setting of the simulation domain boundaries. If the boundary is set as not allowing fluid outflows, the increasing pressure gradient will become reasonable.

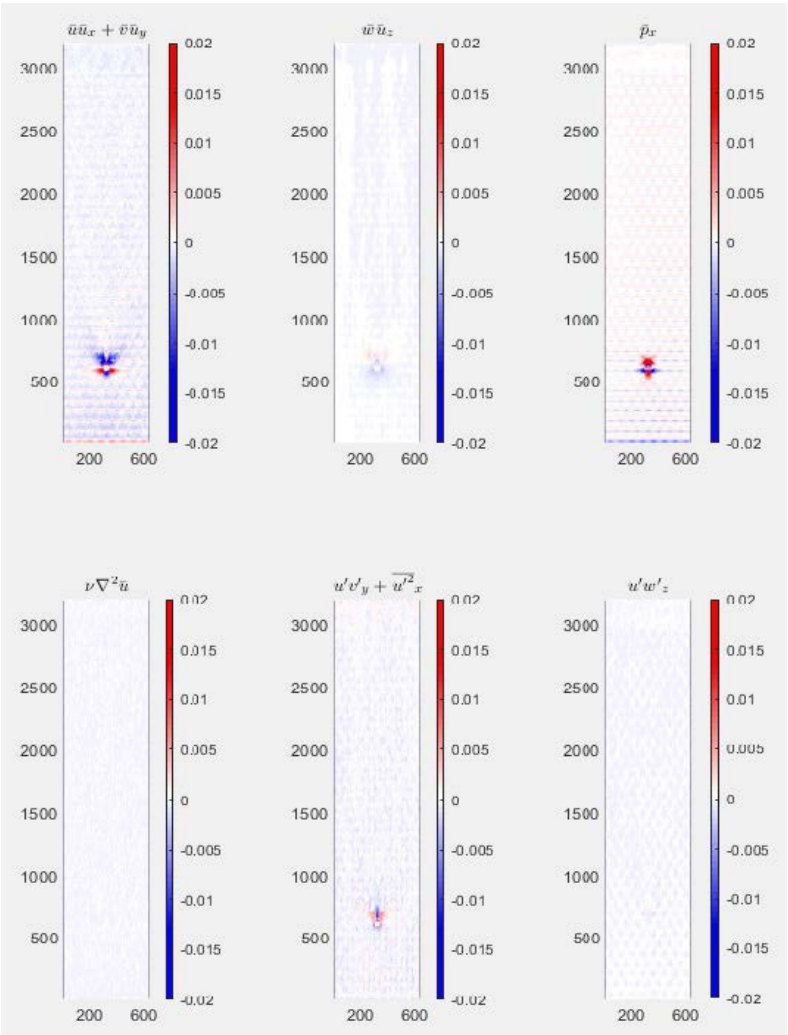

Figure 5. RANS terms at the plane view of the urban area.

Overall, most of the data is reasonable apart from the increasing trend of pressure gradients. Re-running the time-consuming simulation is impossible, not only because the simulation requires a long time but also because running the simulation is not part of this research. Settings for the simulation domain should be checked by the supervisor.

3.2 Gaussian mixture models

GMM are trained on 10% randomly subsampled points for a better clustering speed. The number of clusters for GMM is pre-set as 6 because the governing equation has been divided into 6 sub-parts based on their physical meanings in section 2.1. Setting the cluster number as 6 is reasonable as there will be clusters indicating the same information in the outcomes if the cluster number 6 is larger than the practically required number.

After the model training process, GMM has been applied to the simulation data sets for both side-view and plane-view cases. The covariance matrices are plotted for each identified cluster, as shown in **Figure 6**. The Python configuration is different from MATLAB's, starting the graph from 0 rather than 1; consequently, the cluster numbers in **Figure 6** are from 0 to 5. All matrices are in the form of RANS terms against each other.

Covariance matrices show the correlations of one term to another. Some colours in matrices are shallow while some are not. This is due to the different weights of each term in the momentum balance.

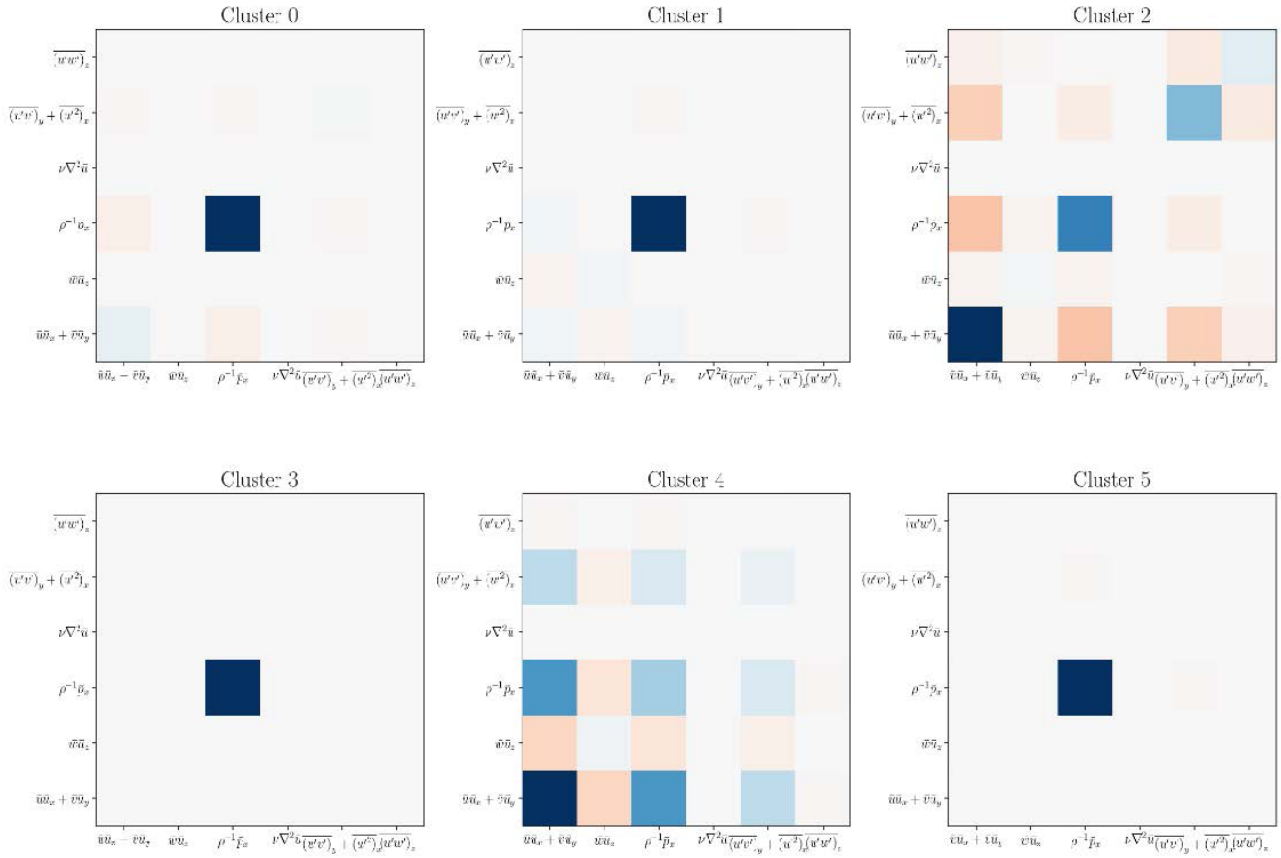

Figure 6. Covariance matrices for 6 GMM clusters of side-view case.

From **Figure 6**, it is obvious that the covariance matrices of cluster 3 and cluster 5 exhibit the same pattern, which only shows the pressure gradient vs. pressure gradient relation; thus, these are trivial results as they do not demonstrate any correlations. Cluster 0 and Cluster 1 are similar but different. In cluster 0, the horizontal advection terms balance the pressure gradient term whilst both horizontal and

vertical advection terms balance the pressure gradient term in cluster 1. Both clusters are reasonable because in the area where the air does not interact with buildings there will only be horizontal advection. Vertical advection exists when flow separation and turbulence occur in the simulation. It is worthwhile to mention that the colour of advection terms in cluster 1 is shallower than that in cluster 0. When the vertical advection term is included for the momentum balance, the weight of horizontal advection terms will be smaller.

In the covariance matrix of cluster 2, horizontal and vertical Reynolds stresses are included in the balance, balancing the horizontal advection with the pressure gradient. The covariance matrix of cluster 4 includes both horizontal and vertical advection, balancing the pressure gradient and horizontal Reynolds stresses.

The same algorithms are applied to the plane-view case. The covariance matrices are also plotted as shown in **Figure 7**. Moreover, the correlations between each term in the plane-view are more complicated than the side-view. The cluster 0 covariance matrix shows correlations between the horizontal advection term, pressure gradient and Reynolds stresses. Horizontal advections are balanced by the pressure gradient and Reynolds stresses. The horizontal Reynolds stress shows strong correlations to horizontal advection term and vertical Reynolds stresses whilst the pressure gradient has a smaller correlation to them. In cluster 1, the relation turns out to exclude the vertical Reynolds stress, instead having a stronger correlation between the pressure gradient and horizontal advection. The pattern of the covariance matrix for cluster 2 is not useful for analysis as it only shows that each term is correlated to itself. Cluster 3 is very similar to cluster 1; the only difference is that the vertical advection is slightly correlated to the pressure gradient in the cluster 3 matrix. The covariance matrix for cluster 4 demonstrates the relationship between advection terms, pressure gradients and horizontal Reynolds stresses. The matrix of cluster 5 is slightly similar to cluster 0 and cluster 3 but with a smaller correlation between Reynolds stresses. In cluster 5, the relation between the horizontal advection term and the pressure gradient is stronger than clusters 0 and 3.

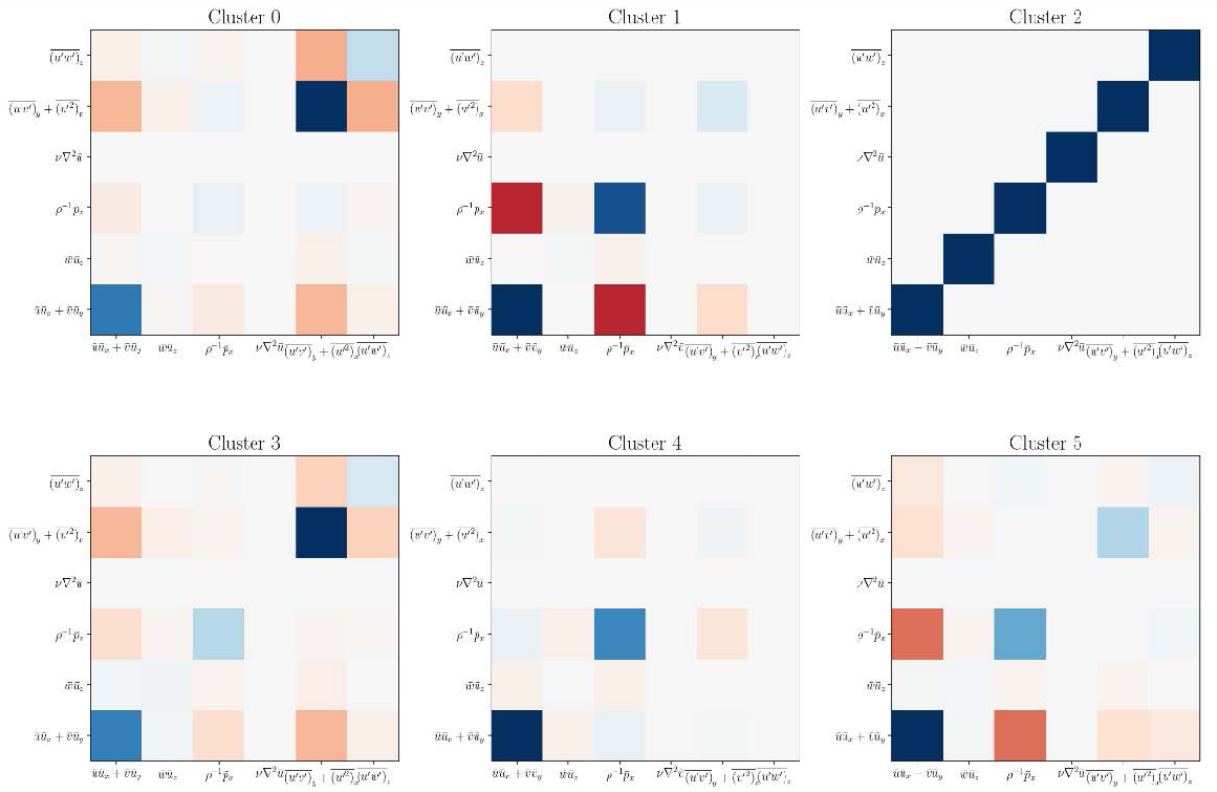

Figure 7. Covariance matrices for 6 GMM clusters of plane-view case.

Overall, the covariance matrices of the side-view case for clusters 0, 1, 2 and 4 are reasonable, which illustrates four kinds of dominant balance regimes. Apart from cluster 2, the matrices of the plane-view case are all useful whilst clusters 0 and 3 are very similar to each other. Five kinds of dominant balance regimes are demonstrated in the plane-view case, with results as discussed in section 3.1. The viscosity term does not contribute to any covariance matrix of both side-view and plane-view cases, because the viscous sub-layer is too tiny in comparison with this macro-scale urban model. Besides, both cases yield results that show trivial clusters exist. To solve this trivial-cluster problem, SPCA is applied, which will be discussed in detail in section 3.3.

Apart from covariance matrices, the scatter diagrams are plotted to visualise GMM clustering with a 2-D view of the equation space. **Figure 8** and **Figure 9** illustrate the 2-D views of equation space for side-view and plane-view cases, respectively.

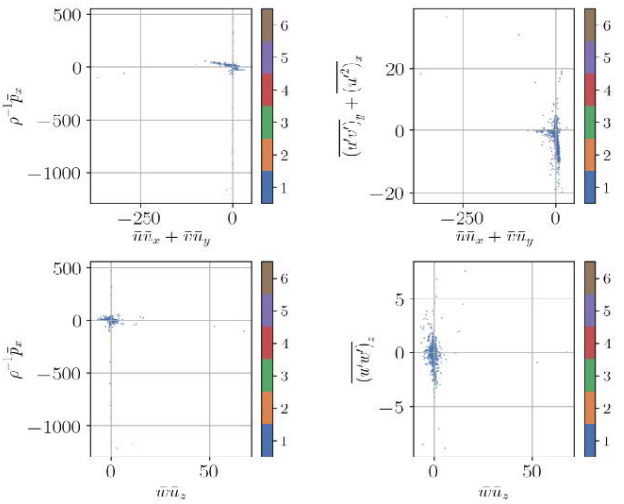

Figure 8. 2-D views of equation space for side-view case.

Six terms in RANS will yield a six-dimensional diagram which is impossible to plot. Therefore, terms are pre-selected based on their correlations in covariance matrices and 2-D scatter diagrams are then generated. In **Figure 8**, four graphs of pressure gradient against horizontal advections, horizontal Reynolds stresses against horizontal advections, pressure gradient against vertical advection and the vertical Reynolds stress against vertical advection are plotted. The graphs are supposed to show different clusters with different colours in terms of scattered points but the diagrams in **Figure 8** can hardly read any clusters apart from the blue cluster which is the first cluster from GMM. As for the top left and bottom left diagrams, there are many points that are far away from the main data points (centred at the origin point (0, 0)). Their corresponding pressure gradient has values of negative hundreds or thousands. The value of **Figure 8** for interpretation and analysis is little whilst **Figure 9** shows a lot more information regarding plane-view data points.

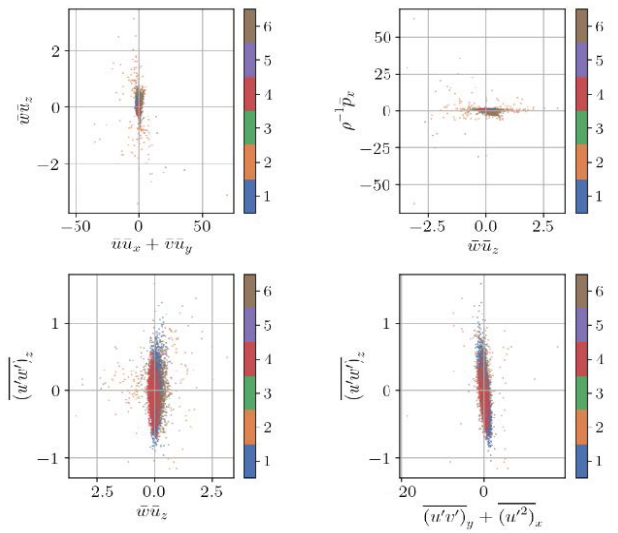

Figure 9. 2-D views of equation space for plane-view case.

Based on the plane-view covariance matrices, scatter graphs containing the relationships between dominant terms are plotted. From left to right, top to bottom, the graphs are of vertical advection against horizontal advections, pressure gradient against vertical advection, vertical Reynolds stress against vertical advection and vertical Reynolds stress against horizontal Reynolds stresses, respectively. Comparing **Figure 9** with **Figure 8,** scatter diagrams in **Figure 9** clearly demonstrate the clusters that each data point belongs to. Most of the points lie in the first, second, fourth and sixth clusters, which is consistent with the results from the covariance matrices in **Figure 7**.

The same GMM algorithms work perfectly on the plane-view urban model, indicating that the monotonous results in the side-view case are not caused by any flaws in the algorithms. There is an enormous range of pressure gradient in **Figure 8** because the variables in

the side-view case are analysed in x and z-directions and the pressure gradient certainly varies with different heights. This also explains the small range in pressure gradient in the plane-view scatter diagrams.

The 2-dimensional GMM clustering results of the entire domain of these two cases are generated. **Figure 10** and **Figure 11** illustrate the side-view case and the plane-view case, respectively.

The clustering results show that most of the regions are in the first and third clusters. The first cluster (blue region) demonstrates the turbulence region. Flow separations and wakes can be clearly identified. Other regions are clustered into the third cluster (green region) where there is no turbulence or very small turbulence. The shape of buildings is clearly shown in the result, apart from the first several small buildings and the skyscraper. The turbulence around the first several buildings would be stronger because the air flows directly to the first building surface and creates a large turbulence area, as shown in the first 500“ in **Figure 10**. Then the air flow meets the skyscraper, which has a much larger interacting surface to the flow than small buildings. Consequently, a much larger wake is formed at the back of the skyscraper. Due to the existence of the skyscraper, the small buildings behind would experience smaller air flows. The shape of the buildings in the clustering result is then clearer than the buildings in the front of the urban area. However, a question may arise: where are the other 4 clusters in **Figure 10** ? The answer is at the front surface of the skyscraper. Having these tiny clusters all lie in the same area is unreasonable. This phenomenon may be caused by the coarse mesh grid of the simulation domain, or by the number of clusters being set larger than was practically required.

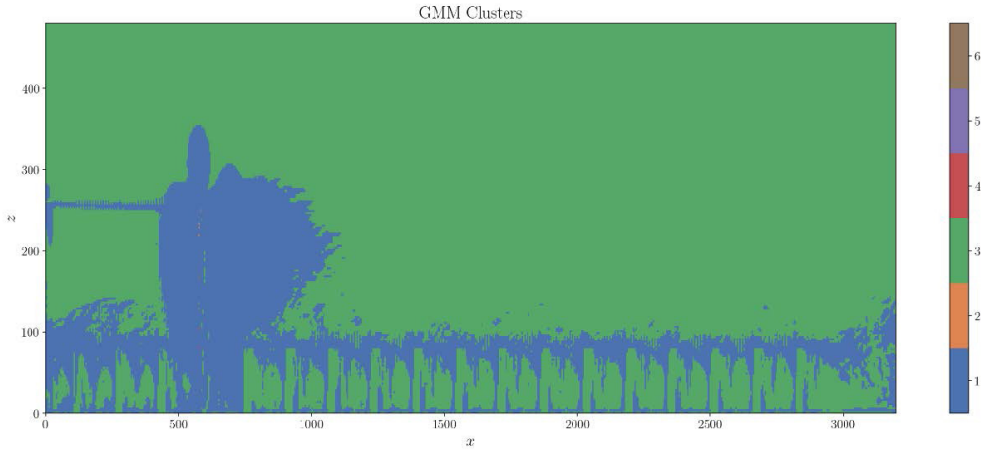

Figure 10. GMM clusters of side-view case.

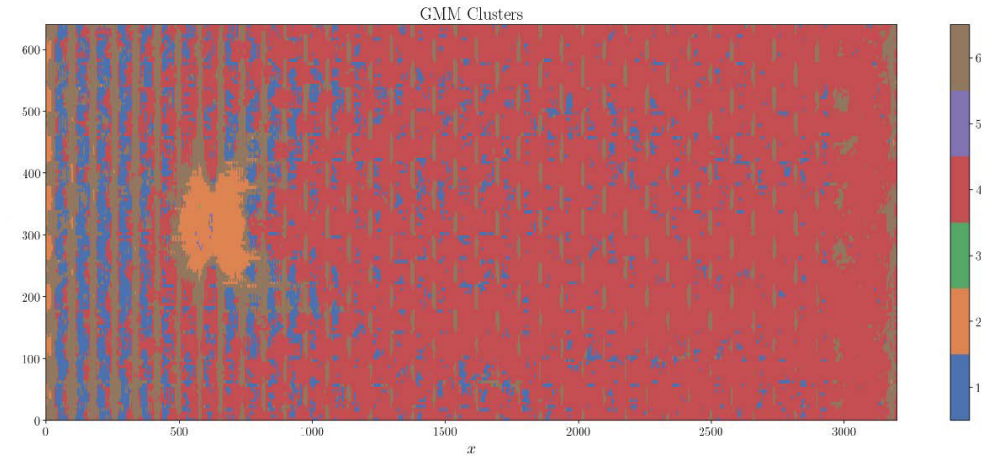

Figure 11. GMM clusters of plane-view case.

The clustering result of the plane-view case in **Figure 11** is more diverse than the side-view case. It is obvious that cluster 2 (yellow region) demonstrates the flow patterns around the skyscraper and most of the region is in cluster 4 (red region). According to **Figure 7,** the first cluster (blue region) has a similar covariance matrix as the fourth cluster (red region) but the first cluster has a larger weight on horizontal advection terms and no weight on vertical advection terms. This indicates that the first cluster focuses on the non-turbulence area whilst the fourth cluster is for the area with some turbulence, which is consistent with the behaviour of these two clusters in **Figure 11** (the blue region mainly on the LHS and the red region mainly on the RHS). Moreover, cluster 6 shows the same pattern as the layout of the urban buildings as shown in **Figure 1** although the small buildings themselves are not included in this slice. Hence cluster 6 contains the region that is under the influence of buildings. Cluster 3 and 5 are located within the region of cluster 2, which is not obvious in **Figure 11**. As discussed in the covariance matrices section, the third cluster is trivial. The fifth cluster seems located around the surface of the skyscraper in **Figure 11**.

The GMM clustering results are justified as reasonable outcomes but there are still excessive trivial clusters existing in both cases. To solve this problem, the SPCA reduction is applied.

3.3 SPCA reduction

The objective of using SPCA is to use l_1 regularisation to extract a sparse approximation to the leading principal component and reconstruct the GMM clusters. As a result, near-zero entries of the leading principal component will be eliminated and active non-zero entries are taken for reconstruction. This process is named SPCA reduction. The model for l_1 regularisation is selected from 10 −4 to 105 and the residual of inactive terms against regularisation graphs are plotted as shown in **Figure 10** and **Figure 11**. In principle, the range of λ can be 0 to positive infinity.

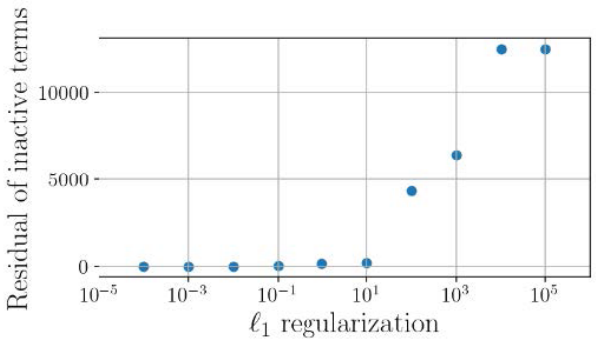

Figure 12. Residual of inactive terms vs. l_1 regularisation for side-view case.

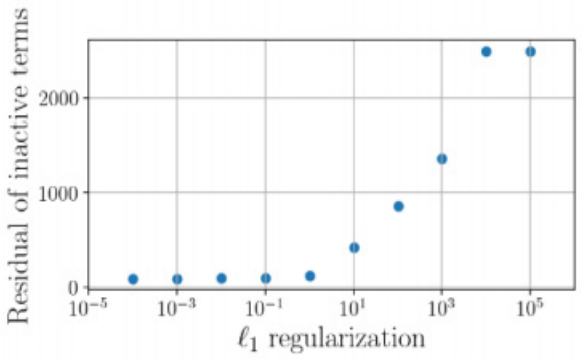

Figure 13. Residual of inactive terms vs. l_1 regularisation for plane-view case.

Figure 12 and **Figure 13** have the same general trend of the residual of inactive terms against the values of λ in l_1 regularisation. The residuals are small before the λ reaches 101 , and it increases significantly when the λ value rises further and reaches its maxima at the values of 104 and 105 . This behaviour matches the expectation of the lasso regression which yields the same answer as the least-square approach when λ is zero and yields a sparse approximation when λ is large enough. Although the patterns are very similar, there is a difference between these two cases. The value of the residuals of inactive terms for the plane-view case is much smaller than for the side-view case. This might be caused by the difference in the GMM inactive term identification. A larger residual indicates that more information in inactive terms is eliminated.

After carrying out the sparse approximation, the graphs of clusters are generated as shown in **Figure 14** and **Figure 15**. The colours for each cluster may have changed but the patterns are the same. Two

tables are generated for dominant term visualisation and the colours used are in the same set of colours as the SPCA reduction graphs. For better interpretation and discussion, two SPCA reduction graphs are put together with the tables in the next section.

3.4 Final dominant balance models

Final dominant balance models can be obtained when the inactive terms are all eliminated during the l_1 regularisation and sparse approximation. Tables showing dominant terms are labelled as **Figure 15** and **Figure 17** accompanied by the SPCA reduction graphs (**Figure 14** and **Figure 16**).

In general, the layout of the clusters is unchanged, but the number of clusters is reduced to 3 rather than 6 from the GMM results. All excessive clusters discussed in section 3.2 are combined into a single cluster (blue) after sparse approximation. However, it is also unreasonable to have a large pressure gradient at the front surface of the skyscraper; no other dominant terms are identified in the blue cluster to balance this pressure gradient. The other two clusters have reasonable dominant balance relations. The green cluster shows a balanced relation of advection terms, pressure gradient and horizontal Reynolds stresses. The orange cluster shows the same relationship but without the vertical advection term because the green region represents the turbulence area whilst the orange region shows laminar or a very small turbulence region. Turbulence generates vortices which leads to a certain value of vertical advection. In the final dominant balance model of the side-view case, the viscous term and vertical Reynolds stresses are not encountered.

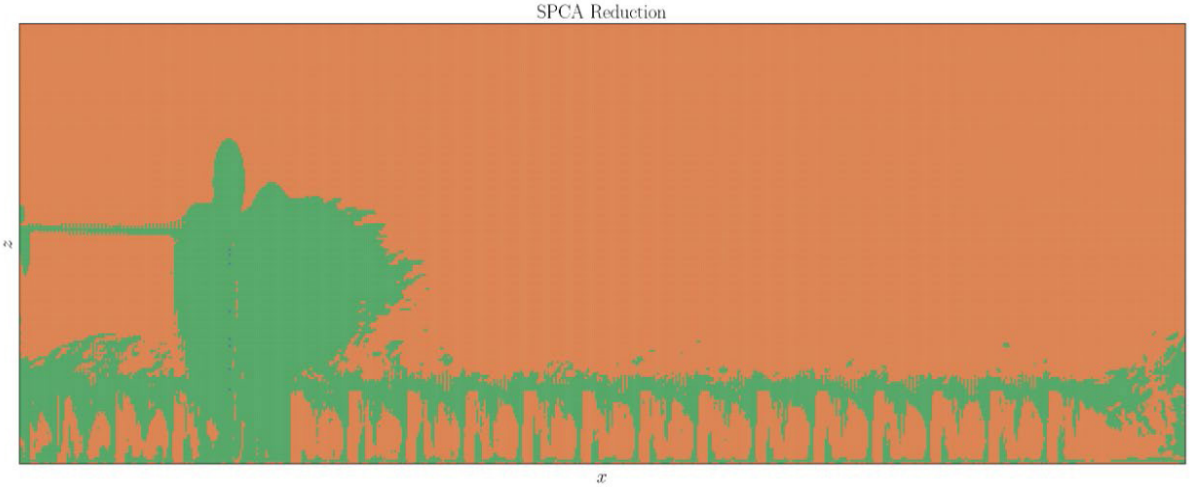

Figure 14. Clusters after SPCA reduction (side-view case).

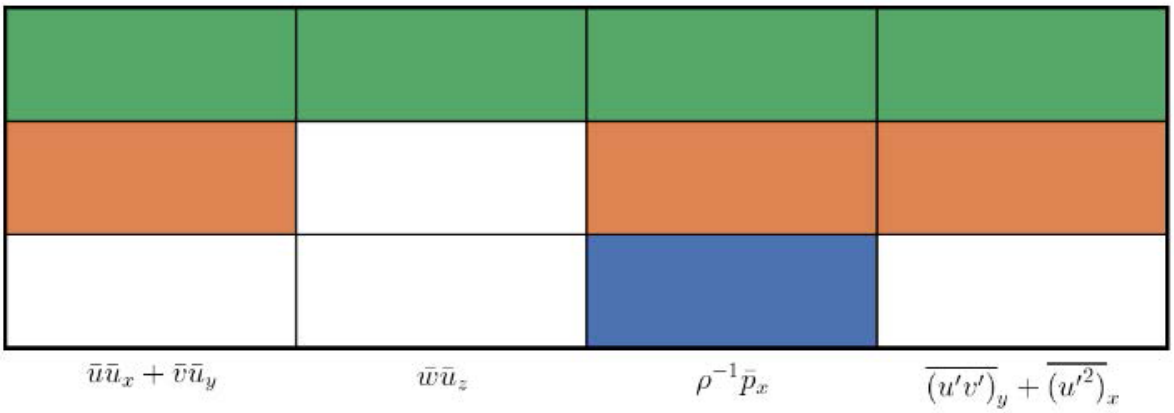

Figure 15. Final dominant balance model for side-view case.

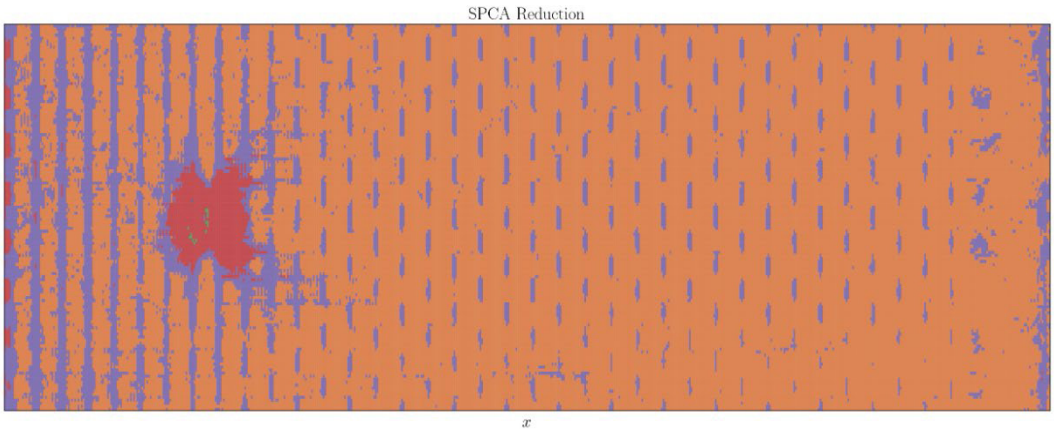

Figure 16. Clusters after SPCA reduction (plane-view case).

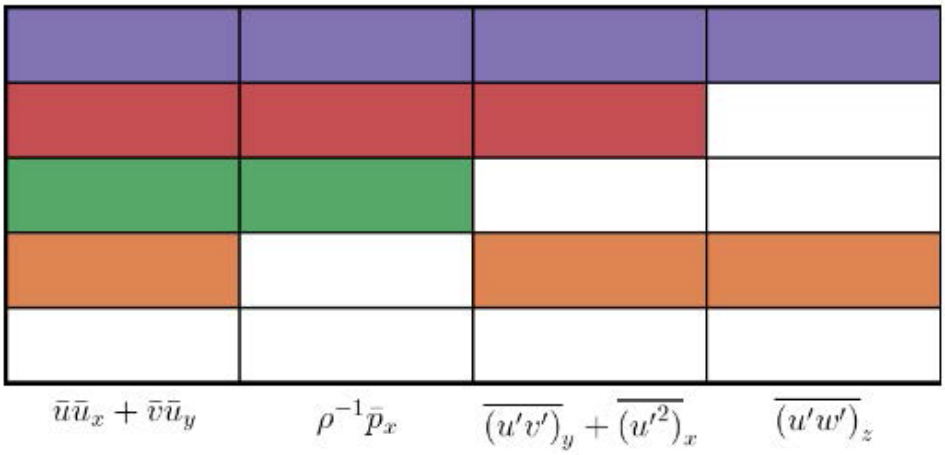

Figure 17. Final dominant balance model for plane-view case.

Comparing **Figure 16** with the previous GMM results (**Figure 9**), a lot of regions at the back of the urban area are approximated into the same orange cluster after SPCA reduction. In this case, the final dominant balance model consists of four clusters with different active terms for balance. Excessive clusters are erased during the cluster reconstruction. The vertical advection and viscous terms are not identified as dominant terms for any clusters, so they are excluded from **Figure 17**. The violet cluster shows the relationship of horizontal advection balancing pressure gradient, horizontal and vertical Reynolds stresses. From **Figure 16**, this cluster is widely evident across the first several buildings and around the skyscraper, showing the high turbulence regions. The red cluster indicates the same balance relation but without the contribution of vertical Reynolds stress. From the graph, the red region is mainly located around the skyscraper because there is no such high building behind it and no vertical vortices can form. Hence the horizontal advection can only cause horizontal vortices which is consistent with the result from **Figure 17**. It should be noticed that there are some regular red clusters on the LHS of the domain. This may be caused by the reflection of the horizontal air flow from the buildings in the first row of the urban model. The green cluster shows a simple relationship between horizontal advection and the pressure gradient at the surface of the skyscraper. The orange cluster shows the dominant balance relation between the horizontal advection and the Reynolds stresses.

The scatter diagrams have been re-plotted after applying the sparse approximation to two cases, which are shown in **Figure 18** and **Figure 19**.

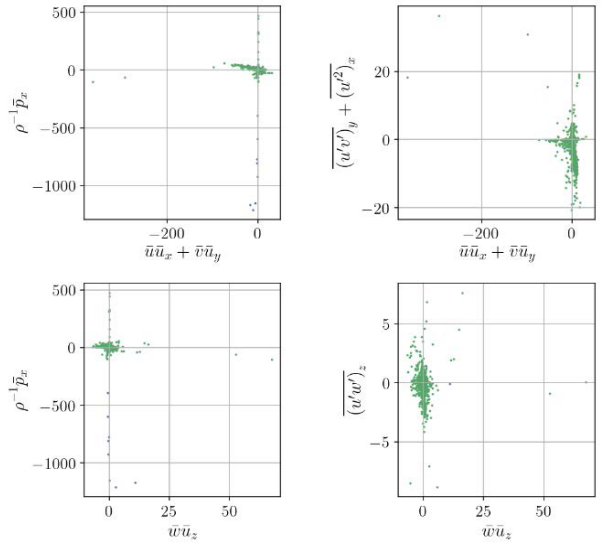

Figure 18. 2-D views of equation space for the side-view case after SPCA reduction.

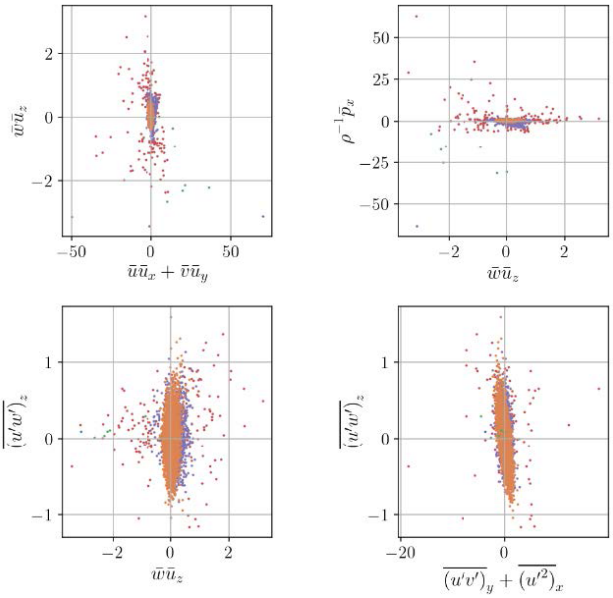

Figure 19. 2-D views of equation space for the plane-view case after SPCA reduction.

In general, the changes in scatter diagrams for both cases after applying SPCA reduction are not obvious, especially for the side-view case. After the sparse approximation, the clusters for the side-view case have been reduced into two normal clusters and one small cluster (pressure gradient). Apart from some blue outliers, all scattered points in **Figure 18** are in green. This pattern is similar to the outcomes in **Figure 8** before SPCA reduction. However, in **Figure 19**, there is an obvious change in the bottom two scatter diagrams. The region of the orange cluster, which was a red cluster in the GMM results, has become larger after the sparse approximation. Some of the noises are eliminated during the reconstruction and forming a larger cluster. The change other than this is tiny, which is hard to observe. More scatter graphs may be worthwhile to be plotted for analysis in further research.

4. Conclusions

In conclusion, Gaussian mixture models (GMM) are efficient in clustering as the model can be trained by only using 10% random samples in a process that takes mere seconds to yield a clustering result of a 640 × 128 or 640 × 192 data set. Taking random numbers for model training is reasonable for GMM because Gaussian distribution is the default setting for random number generation commands. However, there is a drawback to obtaining overlapping or trivial clustering groups because the number of clusters should be pre-defined before applying this to a simulation data set. The solution to these trivial clusters is SPCA reduction, which is covered in this research; this shows a good performance in cluster reconstruction. Moreover, the results from GMM are sensitive to the resolution of the data set. A coarse resolution would lead to inaccurate pattern recognition and clustering. There is a way of dealing with the coarse data set, which is to increase the number of randomly subsampled points for GMM training but the time taken for running the entire script would increase.

Overall, learning dominant urban flows with data-driven models yields reasonable results. Applying machine learning to fluid mechanics can not only shrink the time for post-processing but also presents a brand-new approach to visualising and analysing the physical processes. The essence of the RANS equation is the momentum equation with one term balancing another. Using the dominant balance model can help with understanding the different flow regimes and the dominant terms describing the regime. It is a faster and more efficient way to learn fluid behaviours than the traditional method. This is a general approach that would bring huge con-

tributions to understanding fluid mechanics in any kind of situation, not only for urban fluid mechanics. From the article written by Callaham et al. (2021), the data-driven dominant balance model has been successfully applied to a variety of different physical processes such as the nonlinear optical pulse propagation, geostrophic balance in the Gulf of Mexico and a generalised Hodgkin-Huxley model. Therefore, this method also has great compatibility with important physical processes. It is obvious that the further application of this method in the perspective of fluid mechanics is highly worthwhile and merits further development.

5. Further work recommendations

Due to the limited time for this research, the data-driven dominant balance approach is only applied to the field of urban fluid mechanics. The application in other fields of fluid mechanics is necessary to be carried out to test the approach compatibility in other perspectives.

For this research project, there are some aspects which require further improvement:

- The simulations of the urban model for the neutral case should be re-run with a finer mesh grid to ensure the data set resolution. The resolution of the data set directly affects the quality of the GMM clustering results. For instance, the unreasonable large pressure gradient at the front surface of the high-rise building and the extraordinary turbulence re- gion at the beginning of the domain around 250m could be evitable if a finer mesh grid is applied. Moreover, all graphs showing cluster- ing outcomes suffer from the same resolution problem. This can be observed from the sharp edges between each region in these graphs.
- Running a simulation with a longer time span will improve the clustering results because a longer time span yields more general time-averaged results of the flow field.
- The backward step finite difference method is used for approximating the first-order derivatives in the RANS equation. In comparison to the central difference method, backward step performs a less accurate approximation of the derivatives. Therefore, the backward step method should be replaced by the central difference method to achieve greater accuracy in processing the simulation data sets in further works.
- The Gaussian mixture model (GMM) should be trained by a larger number of random subsampled points for better clustering accuracy although it may take more time for model training and be more demanding of the performance of the computer.

References

[1] UN population Division, 2019. World Population Prospects 2019 Highlights [cited 2024 May 10]. Available from: https://digitallibrary.un.org/record/3813698?v=pdf

[2] Park, Chris C., 2007. A Dictionary of Environment and Conservation (1st Edition). Oxford University Press: Oxford; New York.

[3] Brunton, Steven L., Kutz, J. Nathan, 2019. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press: Cambridge. pp. 21–25, 103–109, 172–176

[4] Callaham, Jared L., Koch, James V., Brunton, Bingni W., et al., 2021. Learning dominant physical processes with data-driven balance models [cited 2024 May 10]. Available from: https://doi.org/10.1038/s41467-021-21331-z

[5] Bishop, Christopher M., 2006. Pattern Recognition and Machine Learning. Springer: New York. pp. 4(4), 738.

[6] Zou, H., Hastie, T., 2005. Regularisation and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 67, 301–320.

[7] Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. Journal of computational and graphical statistics. 15(2), 265–286.