Transforming Data into Meaning. Data-Driven approaches for Particle Physics, Nuclear Power Safety and Humanitarian Crisis Situations.

JOSEPH JAMES WALKER

Transforming Data into Meaning

Data-Driven approaches for Particle Physics, Nuclear Power Safety and Humanitarian Crisis Situations.

Joseph James Walker

A Thesis presented for the degree of
Doctor of Philosophy

Institute for Particle Physics Phenomenology
Department of Physics
Durham University
United Kingdom

December 2023

Transforming Data into Meaning

Data-Driven approaches for Particle Physics, Nuclear Power Safety and Humanitarian Crisis Situations.

Joseph James Walker

Submitted for the degree of Doctor of Philosophy

December 2023

Abstract: Machine learning and data intensive methods can be applied to a plethora of research domains. We apply supervised and unsupervised machine learning, Monte Carlo simulations and statistical tools to three diverse areas of research, tackling a range of computational and data analysis challenges unique to their respective environments.

Using SHERPA – a Monte Carlo event generator – as a Standard Model machine we generate thousands of particle collision events. We employ a range of neural network architectures to determine the most powerful discriminating features which eliminate vast numbers of background events enabling us to calculate new constraints on the charm Yukawa coupling at the Large Hadron Collider and future projections.

Hartlepool Nuclear Power Station has a rich array of instrumentation that continuously monitors reactor health as frequently as every second, at all times. We apply unsupervised machine learning and Bayesian tools to scrutinise anomalous behaviour in the data which is indicative of instrumentation degradation prior to instrumentation failure.

JUNE – an agent based epidemiological simulation – is used to extract novel social

mixing matrices at Cox's Bazar, a refugee camp in Bangladesh containing $\sim 600,000$ displaced people. These contact matrices can be used to understand social interactions and disease spread and therefore provide better utilisation of limited resources.

Contents

Declaration

The work in this thesis is based on research carried out at the Institute for Particle Physics Phenomenology at Durham University. No part of this thesis has been submitted elsewhere for any degree or qualification. This thesis is partly based on joint research and publications noted below:

- Chapter 5 is largely based on "*Constraining the Charm-Yukawa coupling at the Large Hadron Collider*" published in Physics Letters B [1] in collaboration with Frank Krauss.
- Chapter 6 is based on unpublished work in collaboration with Frank Krauss, Richard Bower and Andrew Petts at Hartlepool Power Station, EDF.
- Chapter 7 is based on "*A Mixed-Method Approach to Determining Contact Matrices in the Cox's Bazar Refugee Settlement*" published in Royal Society Open Science [2]. The research of this chapter was conducted in collaboration with Frank Krauss, Difu Shi and Joseph Aylett-Bullock at the United Nations Global Pulse among others.

Acknowledgements

First and foremost, I would like to thank my supervisor Frank Krauss for his guidance and the opportunities he has provided and encouraged me to pursue throughout my PhD journey.

Studying at the IPPP I have been privileged to meet many talented individuals who have become close friends. In particular, thank you to Andrew, Dorian and Parisa, my fellow office mates of OC321 (*let's have some fun*) for some of the best times. Particular highlights over that time included wine time, Pokemon nuzlockes and ballroom dancing. However, being the last one to finish my PhD in that group, the office would have been lonely if not for one more OC321 resident. Peter, I would like to thank you for many beer-fueled conversations, American adventures and wise German proverbs. I would also like to include a special mention to Hitham, Oscar, Peter, Ryan, Sofie and Yunji who very generously agreed to proofread sections of this thesis.

I was fortunate (or forced) to share much of the time over my PhD with my housemate Ben. We got stuck abroad, persevered in lockdown together and survived a questionable quantity of Guinness on many a night. Thank you for the fun times over our time in Durham together.

I would also like to thank my family for their love and encouragement across all my pursuits.

To Kathrine, thank you for your ceaseless encouragement in work and life and the many adventures we've shared. You would think many months of transatlantic

letters would have improved my written English. However, you would be wrong. Thank you in particular for reading through this thesis with highlighter in hand.

The work presented in this thesis has been funded by the Centre for Doctoral Training in Data Intensive Science under UKRI-STFC grant number ST/P006744/1. The industrial work in Chapter 6 was funded by Hartlepool Power Station, EDF through the project "*Investigation into the prediction of plant transients and events utilising novel data analysis techniques*". Lastly, Chapter 7 was funded through the EPSRC IAA project "*Creating humanitarian impact through data modelling: Collaborating with WHO and UN Global Pulse*". The United Nations Global Pulse work is supported by the Governments of Sweden and Canada, and the William and Flora Hewlett Foundation.

A straight line may be the shortest distance between two points, but it is by no means the most interesting

— Doctor Who

Chapter 1

Introduction

Machine learning and data intensive tools can be applied to a plethora of research domains involving high dimensional and complex datasets. These tools have the power to distil meaningful and concise information from large datasets. In this thesis, we will apply supervised and unsupervised machine learning, Monte Carlo simulations and statistical tools to three diverse areas of research. Each of these three distinctive projects tackled a range of computational and data analysis issues which we will address throughout this thesis.

In the first part of this thesis, the introductory chapters will outline the prerequisite concepts required across the projects. We present methods in supervised and unsupervised machine learning in Chapter 2 which are relevant only in Chapter 5 where we use supervised machine learning tools to eliminate backgrounds in a cut-flow analysis and in Chapter 6 where we make use of unsupervised machine learning tools to understand anomalous trends in time-series data. The explanations of the techniques of machine learning and data intensive science are drawn from the reference textbook [3]. In Chapter 3 we will review the stochastic simulation tools employed in Chapter 5 and Chapter 7. In these chapters we make use of SHERPA a particle physics event generator and JUNE an agent based epidemiological simulation tool. In Chapter 4 we will outline the key physics concepts relevant for Chapter 5. In this chapter we largely follow the reference textbooks [4–6].

- In Chapter 5 of this thesis, we will discuss current and future constraints on the measurement of charm Yukawa coupling. Many of the concepts introduced in Chapter 2 and Chapter 3 and Chapter 4 will be utilised in this chapter. We will use the Monte Carlo event generator SHERPA to produce a vast dataset of background and signal processes and apply systemic cuts to preferentially retain signal over background using a selection of neural network architectures. Then the CL_s statistical likelihood profiling method [7, 8] is used to extract a meaningful interpretation of the determined value of the Yukawa coupling, or to be more precise its upper-bound confidence limit.

- In Chapter 6, I present work on the industrial placement undertaken at Hartlepool Power Station in collaboration with Durham University Centre for Doctoral Training in Data Intensive Science and EDF. The placement was conducted between 16th September 2020 – 17th April 2021. The keys aims of this project were to expand upon and explore novel decision support tools which could be used to indicate reactor and or instrumentation health. We apply unsupervised machine learning methods introduced in Chapter 2 to explore the large multi-dimensional time series dataset produced from a large array of sensors in the cooling system and reactor core at Hartlepool power station. We test the applicability of these tools to detect changes in the operational parameters of the instrumentation and the reactor itself. Unexpected or anomalous behaviour can be indicative of instrumentation degradation prior to instrumentation failure. Detection of these behavioural changes allows for safer operation and efficient replacement of key instrumentation. These new monitoring techniques were delivered with a demo dashboard which demonstrates the methods outlined in this chapter for a subset of the operational data measured at Hartlepool power station.

- Lastly, in Chapter 7, I present another industrial placement undertaken with United Nations (UN) Global Pulse in collaboration with Durham University

Centre for Doctoral Training in Data Intensive Science. The placement occurred between 17th January 2022 – 17th August 2022. The UN Global Pulse is a data driven department of the United Nations in which teams of digital engineers work together with global governments to apply modern data science and forecasting tools with novel technologies fulfill the goals of the UN in responding to modern, global challenges [9]. The aim of the placement was to derive new social mixing contact matrices for Cox's Bazar using a combination of survey data and Monte Carlo simulations of the camp. Contact matrices are particularly useful in understanding social mixing and disease spread and as these matrices are not well known for the Global South or vulnerable populations. We develop algorithms within JUNE (an individual based epidemiology tool discussed in Chapter 3) to understand social mixing and derive contact matrices Cox's Bazar, a refugee camp in Bangladesh in which the understanding of social mixing is paramount to limit disease spread and better utilise limited humanitarian aid.

This thesis aims to tie together a range of data driven tools across a wide range of domains of research, pulling together robust tools that are regularly exploited in particle physics to other branches of academic pursuit.

Part I

Introductory Content

Chapter 2

Machine Learning

Machine learning (ML) is an exceedingly powerful tool to aid in the comprehension of vast and complex datasets. It is a rapidly evolving field that is applicable across many areas of research and everyday life [10].

ML is a general catch-all term for algorithms that are able to "learn" without direct supervision to perform a range of tasks requiring some level of pattern recognition. A key benefit of ML tools is that the more data we provide during the training stage the more powerful and accurate they can become, providing the task and dataset are appropriate.

ML models fall into two general classes, supervised and unsupervised. A supervised model requires labeled input and output data during the training processes to inform the learning process whereas an unsupervised model does not and it extracts structures in the data automatically. Any type of supervised neural network requires a carefully curated labeled dataset which can be very resource intensive process to collate. Alternatively, unsupervised methods such as Principal component analysis or K-means clustering do not, they infer arbitrary classifications based on the underlying structure of the data.

ML algorithms can be trained for a number of purposes:

- Classification: An input can be interpreted and classified to a selection of

categories.

- Dimensional reduction: An multi-dimensional input is transformed into a lower dimensional representation with minimal loss of information.
- Anomaly detection: A prediction is made as to how representative a new datum is to the training dataset.
- Prediction: Given an input time series a prediction can be made about the data in the past or future.
- Data generation: The generation of new artificial data which is representative of the training dataset.

An intuitive example of a ML algorithm is the Binary Decision Tree (BDT). They are as powerful as they are computationally quick, and easy to interpret. When trained the BDT classifies the data with a sequence of simple binary decisions, for example *is the petal length* $> 70\,\text{mm}$? [11]. This flow of decisions is often depicted as a branching tree in which every decision creates a new branch.

More complex than the BDT is the Neural Network (NN). NNs are powerful tools which are able to approximate any function that is defined by the data, however as the complexity of the data increases so must the NN and the size of the training set which becomes increasingly challenging to label accurately and efficiently. NNs however can be trained on artificial data (with appropriate considerations), thus we turn our attention to Monte Carlo generated data to train models.

2.1 Machine Learning In Context

In this thesis we will apply ML methods to two areas of research. Firstly, we will apply supervised ML methods to future projections of data produced from the Large Hadron Collider in Chapter 5. Then in Chapter 6 we apply unsupervised ML

methods to the vast array of instrumentation at Hartlepool Power Station which constantly monitor the reactor health.

- In Chapter 5 we turn our attention to the Large Hadron Collider (LHC). The quantity of data produced at the LHC is astronomical. The raw data generated at LHCb from each raw event is roughly 50 kB in size and with as many as 40 million proton beam crossings per second we have 1.5 TB of data being generated every single second. Processing and storing this quantity of data is not economically or technologically possible thus the LHC uses real-time data analysis referred to as "triggering". Triggering is used to discard as much as 99.999% of the data [12]. Despite discarding so much data there is still ∼10PB of data being stored every year which has to be analysed in a smart way. The method of discarding uninteresting events is a cut-flow, this is when a series of successive cuts are applied which will reduce the total number of events in the sample. These cuts can be simple, such as cuts on event properties for example demanding an anti-k_T jet with a minimal p_{T} or they can involve more nuanced relationships derived from complex observables from highly multidimensional datasets. The choice of observables can be determined heuristically, for simple datasets using a good physical intuition. However for increasingly complex datasets machine learning can be a powerful tool. Any cut-flow should preferentially discard background or uninteresting events over the signal or otherwise interesting events. At the LHC BDTs and NNs are particularly popular [13].

  The efficiency and interpretability of BDTs makes them well suited as a trigger level classifier at the LHC. At the LHC they implement a two stage triggering process. The level 1 trigger is automatic and computationally cheap, it looks at surface level features of the events such as very high energy particles or particles in unusual spatial or energetic combinations within the detector. Next, the level 2 trigger takes the information from a whole event and performs more sophisticated analyses and checks for signatures in the data which could be

indicative of new physics.

NNs can be applied to a range of purposes in a high energy physics (HEP) context [14].

- Classification: Extract signal events from a large background or jet classification.
- Dimensional reduction: Creating new and interesting observables, or investigate a dimensional space in which that data has the most variance.
- Anomaly detection: Discover new and interesting physics that are not representative of some model (e.g. the Standard Model).
- Generative Adversarial Neural Networks (GANs): Generate fast event generators.

We will apply a range of NN architectures to create an efficient cut-flow which rejects uninteresting background events preferentially over signal events by many orders of magnitude. This cut-flow will allow us to place tighter constraints on the Yukawa charm coupling using Standard Model Monte Carlo simulated data.

- In Chapter 6 we apply two unsupervised ML methods to provide indicators for transient behaviour at Hartlepool Power Station. The station has a vast array of instrumentation sampling data at a frequency no less than several times per minute. A team of engineers and various tools are already implemented which monitor these readings live at all time to ensure maximal reactor and instrumentation conditions. The safe operation at any nuclear power station is paramount, measuring and understanding every detail of the reaction process is a key part of ensuring the continued safe operation.

  In particular, we examine hourly averaged data over a decade from two reactor cores which have over three hundred fuel rod monitoring channels each with thermocouples and control rod position readouts. Further, each reactor comes

with corresponding cooling systems which monitor: water inlet temperature, outlet temperature, pressure, saline and chemical contents of the vacuum and condenser systems. These readings are deeply complex, containing features sensitive to the power grid demand, ambient temperature, sea temperature and the weather. The features come with complex correlations and periodic behaviour and in some cases missing or corrupted readings. All these challenges have to be addressed before any attempt at applying machine learning methods can be performed.

We investigate two promising avenues for further research which will aid in future tools to monitor reactor and instrumentation health. These two ML methods are complimentary tools which look for anomalies in the data – that is when a set of data points are no longer well represented by underlying distributions understood from previous data. In this project we step away from NNs for the benefit of interpretability, the chosen models provide probabilistic interpretations from easily understood predefined parameters, compared with NNs in which the learnt parameters are not always intuitive.

2.2 Neural Networks

While the structure of a NN is dependent on the form of the input data and the required task the fundamental building blocks remain unchanged. The name and structure of NNs were inspired by the structure and function of the brain, the neurons and synapses in the brain are mimicked in the nodes and interconnected layers in a NN and the firing of neurons is emulated by the activation functions. While this is obviously a charming analogy in fact the complexity, adaptability and functionality of the human brain far exceeds any NN to date. A more accurate picture of the function of a NN is to consider it as learning an approximate function which maps any input data to the output in which the parameters of the function are stored within the structure of the NN itself.

The simplest component of a neural network is the perceptron which consists of an input layer which takes our data, passes it to a node which performs a set of simple operations defined by set of parameters stored in that node and returns an output value. A perceptron is represented graphically in Fig. 2.1.

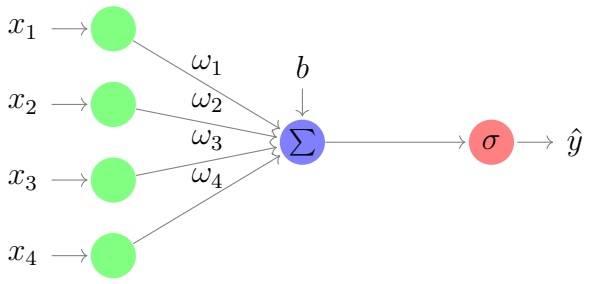


Figure 2.1: The perceptron, the building block of any neural network. A input vector of data x are transformed by the fitted weights ω, bias, b and chosen activation function σ to generate an output value, $\hat{y}$.

The perceptron shown in Fig. 2.1 can alternatively be expressed mathematically as,

$$\hat{y} = \sigma(\sum_{i=1}^{i=4} \omega_i x_i + b) \ , \tag{2.2.1}$$

where x_i defines our input data, ω_i the edge weights, b the bias and finally σ the activation function. Activation functions are introduced to produce a non-linear response and act as a regulator in the NN. There are a plethora of choices and in Fig. 2.2 we show a selection of the most common choices which will become important during the training procedure.

In supervised learning we have a training dataset, X of length N which will have known output values, Y. We can define a loss function which compares the known values Y against the predicted outputs $\hat{Y}$ and informs us how well the NN is performing. The loss function can take a variety of forms; a popular choice is the mean squared error,

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \ . \tag{2.2.2}$$

The performance of the NN can be improved by minimising the loss function with respect to the model parameters – the weights and biases. This is done by a process

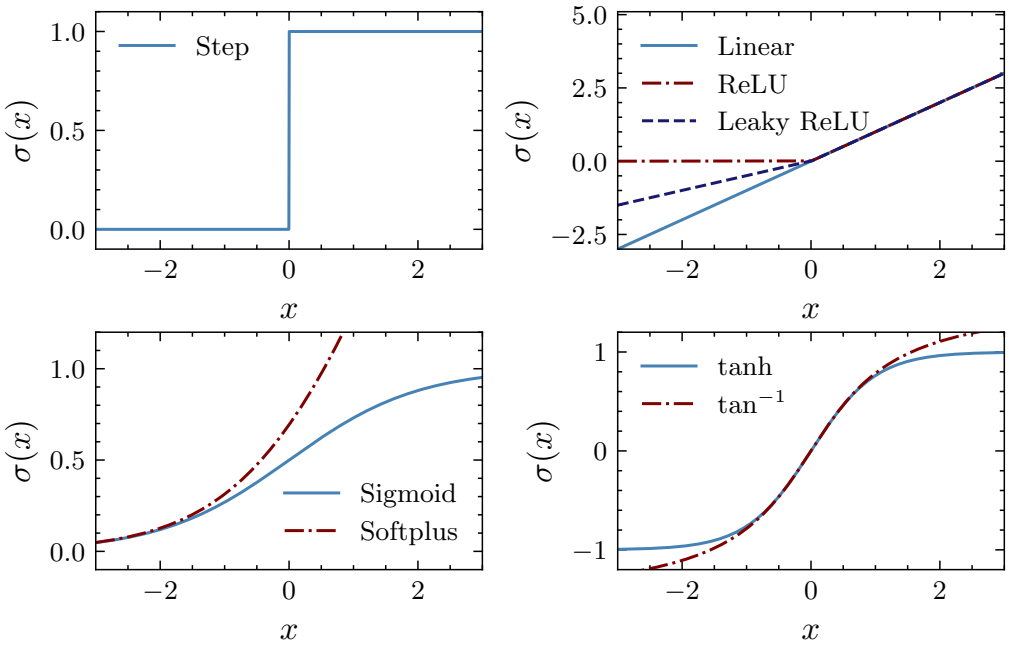


Figure 2.2: A selection of the most common activation functions. We depict a binary step, linear, rectified linear unit (ReLU), leaky ReLU, sigmoid, softplus, tanh and $\tan^{-1}$.

known as backpropagation in which we evaluate the gradient of the loss function with respect to the parameters. This gradient can then be used to incrementally change the value of the parameters in the downwards direction of the loss function landscape at that location in parameter space. A simple numerical method of updating the parameters is a gradient descent approach,

$$\phi' = \phi - \alpha \nabla_\phi L(\phi) \, . \tag{2.2.3}$$

Here ϕ defines any parameter in the model with the primed value being its new updated value, ∇_ϕ is the partial derivative with respect to ϕ and α is the learning rate, a positive scalar that defines the stepsize downhill. The learning rate can be arbitrarily chosen or an optimal value can be selected from a range of values for each iteration of backpropagation – a line search. This process is iterated and produces a downward trajectory in the loss landscape towards – ideally – the global minimum which will be the best set of parameters for the training data and architecture. Gradient descent is a powerful tool for situations in which the minimum of a loss

landscape can not be analytically solved. However, poor choices of the learning rate can yield non-optimal results. If the learning rate is too small, stepsize and therefore convergence can be slow. Further, we may find ourselves in a local minimum in the loss landscape as we have scoped a smaller portion of parameter space. Alternatively, if the learning rate is too large we will jump around the minimum never settling at the minimum. This is why a line search or a learning rate schedule is used where we start with a large value and decrease its value as the loss plateaus towards the minimum.

The training procedure of a NN typically involves splitting the data into (at least[1]) 2 sets, a training set and a validation set. The purpose the training set is to fit the parameters of the model such that the loss function is minimised whereas the validation set is used only for validation and model assessment. Having an independent validation set provides a set of data which the network has never been exposed so that independent performance metrics can be calculated. A NN can be described as overfit if the NN performs extremely well for the training set but poorly for the validation set, this implies that the network is learning the training set precisely and thus generalises poorly on new unseen data. Having an independent check allows us to truncate the training process at an optimal point where the loss is minimised and also when the loss in the validation set is minimised. The training process is outlined below:

1. Forward pass: X_b is passed through the network, calculating the loss function of all data.

2. Backpropagation: The model parameters are updated.

3. Repeat 1. and 2.: Loop through every batch of data, X_b.

4. Repeat 1., 2. and 3.: Loop for a large number of epochs or until some other

[1]More sets can be used for hyper-parameter tuning, model selection or a test set which is carefully curated to contain interesting example data points so that you can compare models with independent data.

criteria is satisfied.

If a training set is especially large it is typical to slice it up into batches to mitigate any hardware bandwidth issues and thus improve overall efficiency. The training procedure and choice of model architecture leaves us with a large number of hyper-parameters to consider. Hyper-parameters define the learning process or model architecture, these include the number of epochs, batch sizes, learning rates and the initial parameter state which directly affect the learning process. We also have the number of nodes, hidden layers, choice of activation functions and loss function which should also be considered carefully as they affect the models' performance for certain tasks and its susceptibility to overtraining. For example, certain activation functions generate a non-linear response which improves the NNs ability to model complex non-linear functions, so we might choose a ReLU or a sigmoid for certain nodes. However, a ReLU can take value 0 which sets its value permanently, it therefore becomes a dead neuron which no longer contributes to the network output. The sigmoid is susceptible to the vanishing gradient problem. This is where the gradients used to update the network become increasingly small approaching zero and precision of the system. This means that the weights of the earlier nodes in the network become slow to update or they stop updating entirely during the backpropagation procedure. Long strings of many interconnected nodes are particularly susceptible to the vanishing gradient problem as the use of the chain rule compounds the effect of the small gradients. These issues can be mitigated by ensuring weights remain close to unity by data normalisation, trying different activation functions or alternatively introducing new structures in large NNs such as recurrent blocks [15] which include skip connections.

Many techniques exist to speed up the training process, make it more generalisable and reduce overfitting:

- Feature engineering: Construct features that are independent and contain high variance.

- Feature regularisation: Normalise and standardise the input data across the sample such that their values typically remain in the interval $[-1, 1]$.

- Data augmentation: Use inherent symmetries in the data to create new data points (e.g. flipping of images) or introducing small changes to the input data to increase network robustness.

- Weight normalisation: The weights of a node can normalised at each step to mitigate the vanishing gradient problem.

- Batch randomisation: Shuffle the data points in each batch at every epoch.

- Dropout layers: Introduce dropout layers which will randomly kill neurons in each batch which serves to reduce overfitting.[1]

With the key concepts of the perceptron defined, we can begin to construct more complex architectures. Firstly, we will discuss dense fully connected neural networks (DFCNs) which are simple feedforward networks composed of stackings of the perceptron in height and width forming new hidden layers. Secondly, we discuss convolutional neural networks (CNNs) which are optimal for datasets which have grid-like topologies. This could be a time-series or images in which neighbouring data points contain correlations which should be exploited and learned by the model by use of convolutional operations. A convolution can be thought of as applying a filter over an image which may extract features such as edges or perform noise reduction with blurring, the parameters in these CNNs learn these types of filter kernels. Lastly, we define recurrent neural networks (RNNs) which process sequential data in a way that parameters are shared over many nodes or edges. They can be configured to take inputs of variable sizes and they contain complex structures which store the state of previous (or future) inputs. Thus they can be powerful tools in interpreting patterns between present and past (or future) data points.

[1] Dropout layers only exist at training time and are removed in validation and deployment.

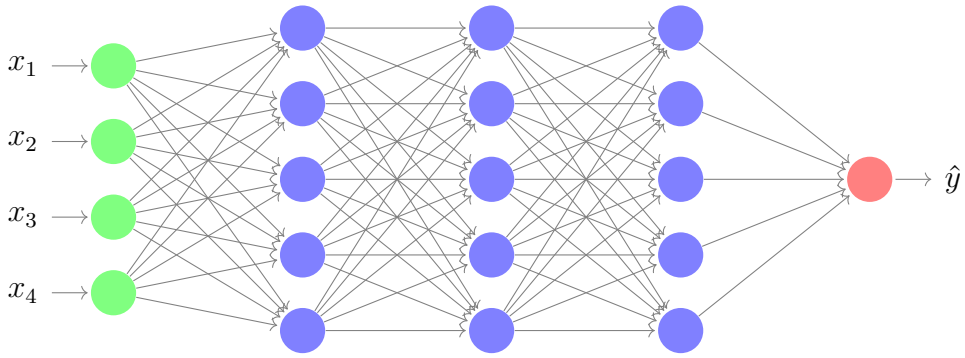


Figure 2.3: An example dense fully connected feed forward neural network. Here a vector $\boldsymbol{x}$ of length 4 feeds into the network which has 3 hidden layers each with depth of 5. A single output value $\hat{y}$ is calculated.

We will use all of these types of networks in Chapter 5 to maximally exploit subtle differences between out signal and background classes:

- Event observables (cf. Sections 4.5.1 and 5.4) → DFCN,
- Jet observables (cf. Sections 4.5.2 and 5.4) → DFCN,
- Jet images (cf. Section 5.4) → CNN,
- Jet particle flows (cf. Sections 3 and 5.4) → RNN,
- Vertex observables (cf. Sections 4.5.3 and 5.4.2) → Time distributed – DFCN,
- Vertex particle flows (cf. Section 5.4.2) → Time distributed – RNN.

2.2.1 Dense Fully Connected Networks

A DFCN is one which comprises of many layers where every node in each layer is connected to every node in the next. A model is described as deep if it contains many hidden layers, this is where the terminology deep learning originates. An example architecture is shown in Fig. 2.3. In a DFCN we can consider each hidden layer as learning increasingly abstracted features.

2.2.2 Convolutional Neural Networks

Convolutional neural networks are powerful tools for data types where the input data is correlated between neighbouring values in some manner. The networks are called convolutional as at least one layer performs a convolution calculation instead of general matrix multiplication, mathematically a convolution is described as,

$$s(t) = (x * k)(t) = \int x(a)k(t-a)da \ . \tag{2.2.4}$$

In ML terminology we describe x as the input and k as the kernel and the output $s(t)$ as the feature map. In the context of CNNs we discretise the input and kernel to handle discrete data,

$$S(i) = (x * k)(i) = \sum_a x(a)k(i-a) \ . \tag{2.2.5}$$

For two dimensional data such as images we can define a two dimensional convolution,

$$S(i,j) = (x * k)(i,j) = \sum_a \sum_b x(a,b)k(i-a, j-b) \ . \tag{2.2.6}$$

A two dimensional convolution can be thought of as passing a kernel over an image and summing element-wise the products of the elements that overlap the kernel as depicted in Fig. 2.4. The convolution can of course be extended to an arbitrarily large number of dimensions. It could be suggested that one could use a DFCN network for image processing, an array describing an image could be flattened into a one dimensional array and training could commence. However, the properties of CNN architectures make it well suited for image processing as they have increased performance with reduced complexity. A CNN is sparsely connected, the nodes of each layer are not connected to every node of the previous layer which does two things: the number of calculations is greatly reduced and the receptive field of the nodes are reduced. The receptive field is the collection of nodes that affect each other, a reduced receptive field (in the context of kernel filters) forces the early hidden layers to detect smaller meaningful features like edges first before the more abstracted features spanning whole images. Secondly, we have parameter sharing,

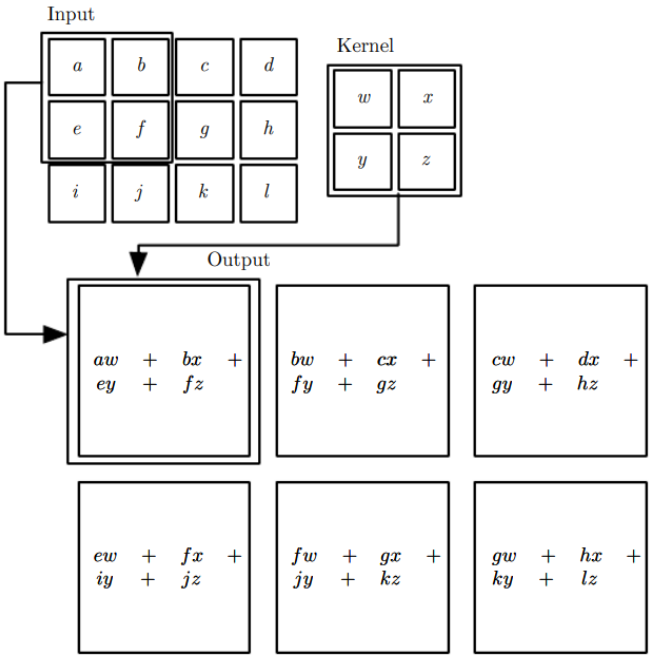


Figure 2.4: An algebraic example of two dimensional convolution. This convolution is "valid", the convolution is only calculated when the kernel overlaps the image fully. Figure from [3].

each convolutional layer defines a kernel which acts over many nodes reducing the complexity of the neural network substantially. Using a DFCN for a purpose well suited to a CNN is likely to be susceptible to poorer performance, longer training times and overfitting. A convolutional layer is defined by three steps:

1. Convolution stage,
2. Detector stage,
3. Pooling stage.

The convolution stage we have already discussed it is the stage which performs the convolution of nodes with a kernel. The detector stage applies activation functions such as those defined in Fig. 2.2 to the node outputs for the same purposes as the DFCN – to introduce non-linearity. Lastly, the pooling stage calculates a summary statistic over a region in the image, for example the average, minimum or maximum value. Pooling is introduced to enforce translational invariance, we generally wish

to be sensitive to the feature existing but not to exactly where it exists in the image. Pooling also reduces the dimensionality of the data flowing through the network reducing the complexity of the CNN. After a handful of convolutional layers their outputs can be combined in the architecture to transition into fully connected hidden layers and finally to the output layer.

2.2.3 Recurrent Neural Networks

Recurrent neural networks are used to process one dimensional sequential data such as speech, written language or dynamical systems. There are two key differences between the RNNs and DFCNs: an RNN takes advantage of parameter sharing like the CNN but also an RNN is "stateful". A stateful network is one which retains some information about a past or future state in the data, enforcing connections between difference features in a time-correlated way. To define the structure of a RNN we consider a classical dynamically evolving system which is defined by some function, f with parameters, θ that maps the previous state, $s^{(t-1)}$ to the new one $s^{(t)}$,

$$s^{(t)} = f(s^{(t-1)}; \theta) \ . \tag{2.2.7}$$

In the context of RNNs we can instead consider a hidden layer dynamical state, h that depends on some driving force, x the features in the data input vector,

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta) \ . \tag{2.2.8}$$

This approach is particularly powerful as it allows us to iterate arbitrarily many times thus we can feed in input data of arbitrary length.

There exist a few choices for how to use the state information, it can be passed only to other RNN cells in the past or future or alternatively also to other layers in the network (cf. Fig. 2.5 depicts the state being passed in to other layers and future RNN cells). Each of these options comes with a unique balancing act between complexity and training efficiency. Simple RNNs (see Fig. 2.5) have significant draw

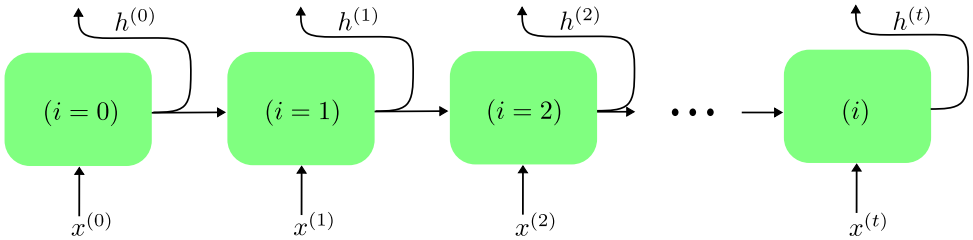


Figure 2.5: An example neural network consisting of a chain of RNN cells. The state of each cell, $h(t)$ is fed into the next cell and can be fed to further layers deeper in the model.

backs: they are highly susceptible to the vanishing gradient problem and they have a short-term memory in which they have a hard time retaining information over long sequences. To combat these issues more complex RNN layers: the Long Short Term Memory (LSTM) and the Gated Recurrent Unit (GRU) layers were invented. Both of these comprise of cells which introduce a set of structures known as gates which calculate thresholds in the data indicating a level of retention or importance. In this thesis, we will make use of GRUs only. In each GRU cell i, we define a set of input weights $\omega_{i,j}$, recurrent weights $\psi_{i,j}$ and biases b_i which parameterise the activation of the reset and update gates and the hidden layer. u and r correspond to the update and reset gates activation,

$$u_i^{(t)} = \sigma(b_i^u + \sum_j \omega_{i,j}^u x_j^{(t)} + \sum_j \psi_{i,j}^u h_j^{(t)}) \,, \tag{2.2.9}$$

and

$$r_i^{(t)} = \sigma(b_i^r + \sum_j \omega_{i,j}^r x_j^{(t)} + \sum_j \psi_{i,j}^r h_j^{(t)}) \,. \tag{2.2.10}$$

Finally we can define the update equation for the state,

$$h_i^{(t)} = u_i^{(t-1)} h_i^{(t-1)} + (1 - u_i^{(t-1)}) \tanh\left(b_i + \sum_j \omega_{i,j} x_j^{(t)} + \sum_j \psi_{i,j} r_j^{(t-1)} h_j^{(t-1)}\right) \,, \tag{2.2.11}$$

where $\boldsymbol{x}^{(t)}$ is the input vector of features for a time t and $\boldsymbol{h}^{(t)}$ is the current hidden layer state vector. Sigmoid, σ and tanh are the choice of activation functions used in the GRU such that the update and reset values are regulated between 0 and 1

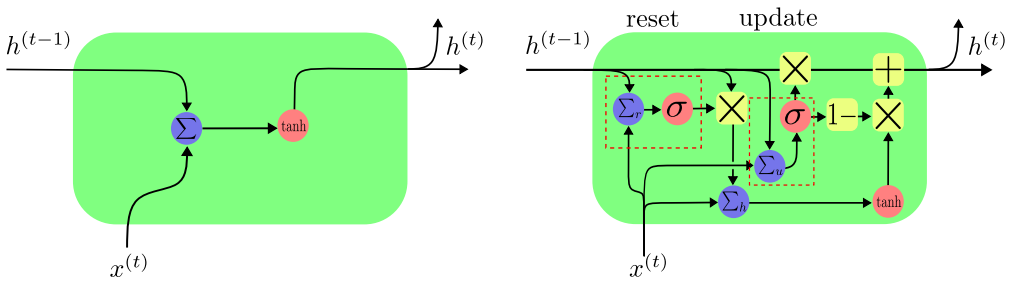


Figure 2.6: Left: A simple RNN which takes information from the hidden state of the previous cell to determine its output. Right: A GRU, introducing an update and reset gate indicated by the red dashed bounding box. $\sum$ represents the weighted sum plus bias, $+$ and $\times$ are element wise addition and multiplication and $1-$ transforms an input x as $x \to 1 - x$. Finally σ and tanh represent the sigmoid and tanh activation functions.

and the overall cell activation between -1 and 1. This keeps the cell regulated and allows the update and reset gates to individually ignore features or mark importance in the state vector.

Finally, we introduce time distributed layers which slice an input vector and apply the same identical NN layer across each slice. This is a useful tool if we want each slice of the data to undergo the same transformations from a given layer in the neural network, thus extracting the same features from each. The outputs from the time distribution can then be concatenated into dense layers.

2.3 Unsupervised Machine Learning

We have discussed at length neural networks which are an examples of supervised machine learning, in Chapter 6 we will make use of two examples of unsupervised learning, Principal Component Analysis (PCA) and Bayesian Online Change Point Detection (BOCPD) which are tools we employ for anomaly detection.

2.3.1 Principal Component Analysis

PCA [3] is a powerful unsupervised machine learning tool which transforms a high dimensional dataset into a lower one by transforming the data along components containing the most variance $1, 2, ..., k$ and dropping the lower variance dimensions $k+1, k+2, ..., m$. Dimensional reduction allows for easier data visualization, quicker computation and easier model training without loss of information. If we keep k components containing a threshold variance we can ensure information about the features is not lost. Intuitively we define the 1^{st} PCA component direction with the unit vector $\hat{\boldsymbol{w}}$ as such:

$$\sigma_{k=1} = \text{argmax}\{\sum_i (\boldsymbol{x}_{(i)} \cdot \hat{\boldsymbol{w}})^2\} \ . \tag{2.3.1}$$

Alternatively the PCA components can be determined by calculating the eigenvectors $\boldsymbol{v}_k$ and eigenvalues λ_k of the matrix $\boldsymbol{x}^{\text{T}} \cdot \boldsymbol{x}$ where $\boldsymbol{x}$ is our $n \times m$ matrix dataset with n rows of samples with m columns of features. The eigenvectors form the weight matrix $\boldsymbol{W} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_m)$ which are ordered in terms of their eigenvalues such that $\lambda_i > \lambda_{i+1}$ therefore generate a diagonalised matrix $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \lambda_2, ..., \lambda_m)$. Then the transformed data is;

$$\boldsymbol{P} = \boldsymbol{W}^{\text{T}} \cdot \boldsymbol{x} \ . \tag{2.3.2}$$

Much like the supervised neural networks, PCA performs best on standardised and centred values which are calculated with the following transformations;

$$x_{i,j} = \frac{x_{i,j} - \mu(\boldsymbol{x}_j)}{\sigma(\boldsymbol{x}_j)} \ , \tag{2.3.3}$$

where $\mu(\boldsymbol{x}_j)$ and $\sigma(\boldsymbol{x}_j)$ are the mean and standard deviation across all available samples for feature j. Nominal data to one standard deviation will now exist in the range $[-1, 1]$.

PCA works under the assumption that the measurements are independent and stationary. For a non-stationary series, any new data point will be poorly represented in the transformation thus we can use PCA as a means of anomaly detection. Further,

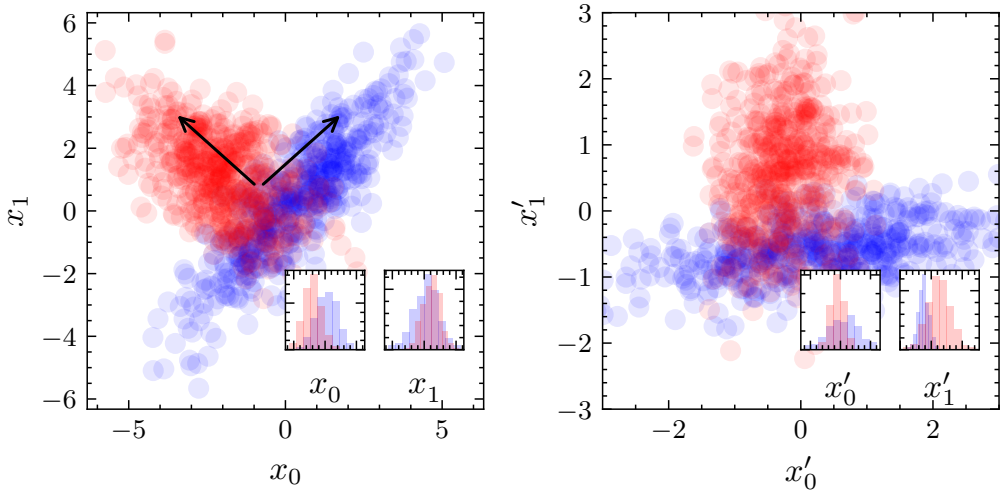


Figure 2.7: Operational example of PCA on a dummy dataset. Left: Depiction the raw data and arrows indicating the PCA component directions. Right: Normalised and transformed data. Histograms over each the axis are also shown.

if new data points in general are poorly represented then this might be an indication of changing distributions in the original, un-transformed data. To quantify this we introduce two test statistics, the Hotelling-T^2 statistic and the square predictive error also known as the Q statistic. These statistics are often used to monitor time dependent processes in various industrial settings [16–18] not dissimilar from the data in Chapter 6. The T^2 statistic is defined as;

$$T^2 = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P}_{:k} \boldsymbol{\Lambda}_{:k}^{-1} \boldsymbol{P}_{:k}^{\mathrm{T}} \boldsymbol{x} \ . \tag{2.3.4}$$

Which gives us a measure of deviations of the latent variables as understood within the first k PCA variables. If we have large deviations in these latent variables we see an increase in T^2. A constraint can be placed on the values of T^2 which we deem an acceptable variation in our data with a confidence limit,

$$\alpha_{T^2} = \frac{k(n^2-1)}{n(n-k)} F_{k,n-k}(\alpha_{\mathrm{crit}}) \ . \tag{2.3.5}$$

$F_{k,n-k}(\alpha_{\mathrm{crit}})$ defines the Fisher-Snedecor distribution with a confidence α_{crit} which for the 99% confidence limit, $\alpha_{\mathrm{crit}} = 0.01$. The Fisher-Snedecor distribution is a

continuous probability distribution that is often used as a tool in analysis of variance or F tests [19]. The Q statistic is defined as such;

$$Q = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{:k}\boldsymbol{P}_{:k}^{\mathrm{T}})\boldsymbol{x} \ . \tag{2.3.6}$$

The Q statistic gives us a measure of the goodness of fit of the sample and is directly related to the models understanding of the noise of the $m-k$ rejected PCA variables. The Q statistic can also constrain our confidence in the sample data point with the approximation [20],

$$\alpha_Q = \theta_1 \left(\frac{z_{\alpha_{\mathrm{crit}}}\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0(1-h_0)}{\theta_1^2} \right)^{1/h_0} , \tag{2.3.7}$$

where e.g. $z_{\alpha_{\mathrm{crit}}} = 2.33$ is the z score required for $\alpha_{\mathrm{crit}} = 0.01$ in a two tailed confidence limit and

$$\theta_i = \sum_{j=k+1}^{m} \lambda_j^i \quad \text{where} \quad i = 1,2,3 \quad \text{and} \quad h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \ . \tag{2.3.8}$$

Comparing the PCA transformed values, T^2 and Q statistic enables PCA to be used as a form of anomaly detection. Any values of large magnitude in the transformed space or in these statistics then indicates that the data point is badly explained by the PCA model.

2.3.2 Bayesian Online Change Point Detection

Bayesian modelling is broadly speaking any modelling technique that incorporates a degree of certainty into the model building. Here we define a Bayesian approach to detecting "change points" (CPs). A change point is described as a point in the data at which we see a distinct change in the properties of the underlining distributions producing that data. There are a plethora of CPs detection algorithms [21], however a key benefit of a Bayesian approach is the probabilistic nature of the method and interpretability it provides.

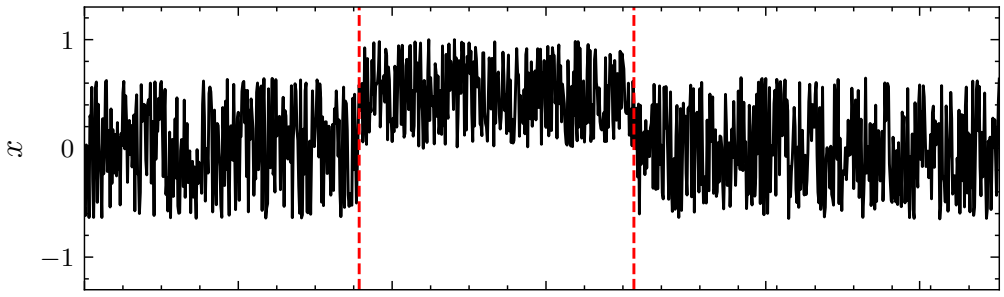


Figure 2.8: An example piecewise distribution with two change points located at the red dashed lines.

Consider the example shown in Fig. 2.8 which depicts a piecewise distribution made from three Gaussian distributions which have different parameters in each of the three regions. The point at which these underlying parameters (the mean and variance) change are known as change points. The goal of Bayesian change point detection (BCPD) is to identify where these CPs lie with some confidence which depends on our priors the best estimate for the underlying distributions and the uncertainty on the data and the data itself. BCPD falls into one of two types, online and offline. Offline requires the full dataset before looking for CPs which typically means our priors are better informed but we lose the possibility of live forecasting. Bayesian online change point detection (BOCPD) allows us to actively update our priors as we collect more data and return probabilistic certainties of CPs. We follow the methodology of Adams and Mackay [22], making only minor adjustments in how the sequence parameters are handled.

We describe CPs in data in terms of run lengths l, for every additional data point if we consider it part of the same distribution $l_{t+1} = l_t + 1$ i.e. the run length has increased by one. Alternatively, if we consider the new data point belonging to a different distribution then $l_{t+1} = 0$. To take a probabilistic approach, we consider all possible run lengths, for example a data point at $t = 5$ has 6 possible run length paths (see Fig. 2.9).

New data points continue as a member of the sequence of length $l_t = t - t_{start}$ of all previous points, which start at t_{start}. A drop to $l_t = 0$ indicates the current run

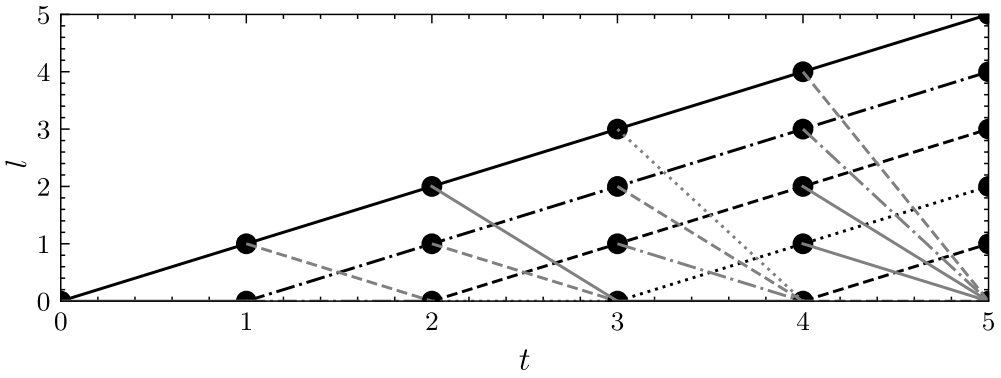


Figure 2.9: Example of all possible run length paths between $t = 0$ and $t = 5$.

is truncated and a new sequence begins – there is a CP. The BOCPD algorithm gives each (t, l) coordinate a probability value which is normalised to 1 over t of constant value, these probabilities are mapped to a matrix $\boldsymbol{L}(t, l)$ which identifies the most likely t where the underlying distributions have changed. To implement this procedure, we first have to make an underlying assumption about the distribution our data originates from, our prior $\pi_t^{(l)}$. This prior depends on the distribution of the data in all possible possible run lengths l_t at time t. It is common to assume that the data is approximately drawn from a Gaussian distribution with $\mu(\boldsymbol{x})$ indicating the mean of the series $\boldsymbol{x}$ and α is standard error such that,

$$\pi_t^{(l)} = \mathcal{N}\bigg(x_t \quad | \quad \mu = \mu(x_{t-l} : x_t), \sigma = \max\Big(\alpha_{t-l} : \alpha_t\Big)\bigg) \,. \tag{2.3.9}$$

Next we define a hazard function H this is a hyper-parameter estimator of the frequency of CPs expected in the series, we can define it as,

$$H(\tau) = \frac{1}{\lambda} \,, \tag{2.3.10}$$

where λ is a timescale estimator for a memoryless process where we consider CPs to be independent. The analysis is not hugely sensitive to this value if the CPs are clearly defined or the errors are narrow with respect to the magnitude of the data. The algorithm is as outlined in Alg. 1.

Alg. 1: The BOCPD algorithm.

Input **Data:** $\boldsymbol{L}(t,l) = 0$, except $\boldsymbol{L}(0,0) = 1$
t: The active time step
l: The sequence length
if Online **then**
  Observe new datum: x_{t+1}
  for $l \in \mathit{length}(x_t)$ **do**
    Calculate mean of sequence: $\mu_t^l = \mu(x_{t-l} : x_t)$
    Calculate estimated standard deviation of sequence: $\sigma_t^l = \max\Big(\alpha_{t-l} : \alpha_t\Big)$
    Determine priors: $\pi_{t+1}^{(l)} = \mathcal{N}(x_{t+1}|\mu_t^l, \sigma_t^l)$
    Calculate growth probabilities: $\boldsymbol{L}(t+1, r+1) = \boldsymbol{L}(t,l)\pi_{t+1}^{(l)}(1-H)$
  end
  Calculate CP probabilities: $\boldsymbol{L}(t+1,0) = \sum_{l_i=1}^{l} \boldsymbol{L}(t,l_i)\pi_{t+1}^{(l)}H$
  for $l \in \mathit{length}(x_t)$ **do**
    Determine sequence length probabilities: $P(l) = \dfrac{\boldsymbol{L}(t+1,l)}{\sum_{l_i=0}\boldsymbol{L}(t+1,l_i)}$
    Update growth probabilities: $\boldsymbol{L}(t+1,l) = P(l)$
  end
end
Set new t : $t = t+1$

In this algorithm an additional optimization can be used to improve computational efficiency, we truncate the series from l_{crit} and above given $\sum_{l_{\text{crit}}}^{t} \boldsymbol{L}(t,l) < \beta$ for a given threshold probability β. Here we take $\beta = 10^{-5}$. This prevents a long probability tail for large l and unnecessary computation on vanishingly small probability sequences.

Running the BCPD algorithm on the test dataset (Fig. 2.8) we produce probability matrix $\boldsymbol{L}(t,l)$ depicted in Fig. 2.10. We see drops from high sequence lengths l to low values around where the CPs are known to exist. It should be noted BOCPD works only on de-trended or stationary data. A stationary dataset is one in which the mean and variance do not change over time. If there is any seasonality in the data it will interpret this as an almost continuous set of CPs because the underlying distributions are continuously evolving.

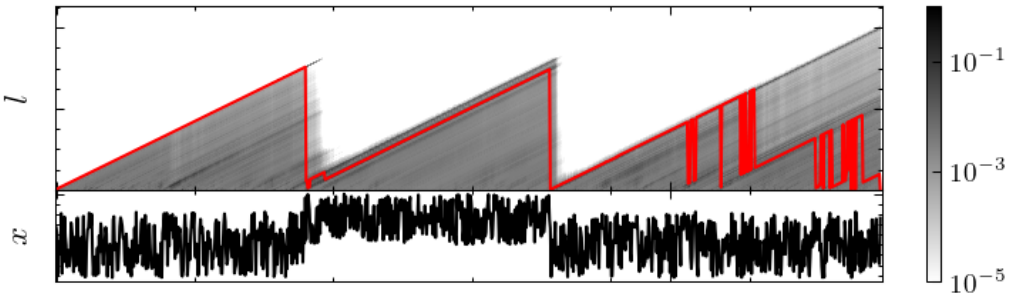


Figure 2.10: Top: Colour map of matrix $\boldsymbol{L}(t, l)$ with red indicated l_{max} where the probability is largest for each time step t. Bottom: An example piecewise Gaussian data like Fig. 2.8.

2.4 Performance Metrics

Throughout this thesis, we use ML tools to perform classification tasks where we wish to filter data points into predetermined categories. To determine the effectiveness of our models it is common to examine a confusion matrix and its associated elements. A confusion matrix $\boldsymbol{\mathcal{C}}$ tells us how data points are classified by our model splitting

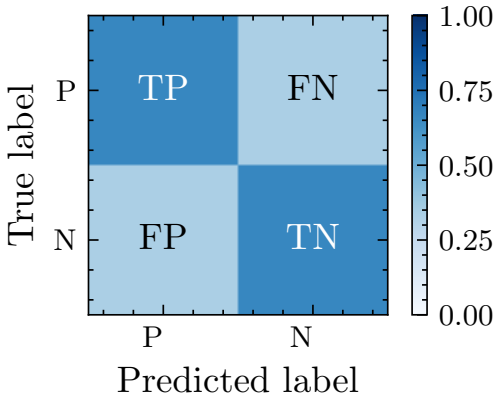


Figure 2.11: An example confusion matrix in which data points of two classes P and N can be classified into, true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

up the results into true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) and we use these to define other metrics. The confusion matrices can be normalised $\hat{\boldsymbol{\mathcal{C}}}$ across their rows to present percentage classification

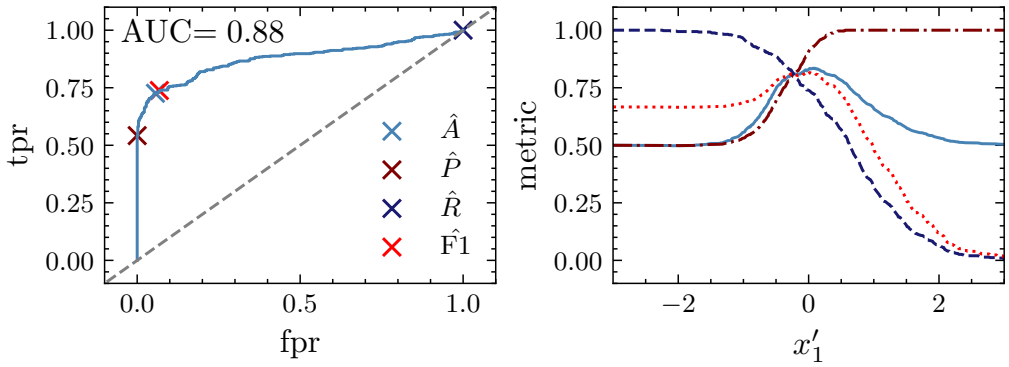


Figure 2.12: Left: Example ROC Curve from data depicted in Fig. 2.7. Key threshold choices maximising each of accuracy A, precision P, recall R and F1 score are marked. Right: Metrics A, P, R and F1 as a function of the the decision boundary x_1'.

rates into the predicted classes. Accuracy A is defined as,

$$A = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} , \tag{2.4.1}$$

which tells us how well we correctly predict data points. However this is not necessarily the best metric depending on the task in hand. Accuracy is a poor metric if we have unbalanced classes or we care about the relative ratios of TP to FN or FP. It is worth considering the precision P and recall R metrics,

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} , \tag{2.4.2}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} , \tag{2.4.3}$$

P and R can be weighted together to create the F1 score,

$$\text{F1} = \frac{2PR}{P + R} , \tag{2.4.4}$$

which is a balance between the precision and recall. Receiver Operating Characteristic (ROC) curves are a powerful way to assess the classifying power of a model, each point on a ROC curve corresponds to a threshold which leads to a particular

false positive rate (fpr) and a true positive rate (tpr),

$$\text{fpr} = \frac{\text{FP}}{\text{FP} + \text{TN}} \,, \tag{2.4.5}$$

$$\text{tpr} = \frac{\text{TP}}{\text{TP} + \text{FN}} \,. \tag{2.4.6}$$

The threshold can be selected to maximise our chosen metric (accuracy, recall, precision or F1). In Fig. 2.12 we show the ROC curve produced from data depicted in Fig. 2.7 where we build a simple classifier using x_1' as the decision boundary. The closer the ROC curve pushes towards the top left the better the classifier is performing, if the ROC curve is a diagonal line the classifier is a random guesser these are characterised by the area under the curve (AUC). For a binary classifier an AUC of 1 is a perfect classifier, 0.5 a random guesser and 0 an anti-classifier (i.e. inverted labels).

Chapter 3

Monte Carlo Simulations

Monte Carlo simulations are computational techniques in which we aim to produce data representative of reality using stochastic methods. They are particularly powerful tools as they can be used to produce artificial data which is based on a series of priors and model assumptions. The artificial data can be used to scrutinise our underlying assumptions or theoretical models used to originally create it. Thus, we have a way to independently check our understandings of the physical word in a quantitative and statistical way.

Monte Carlo methods are particularly well established in particle physics for the modelling of high energy collisions. The behaviour of particles can be understood to follow stochastic processes defined by our theories. The behaviour of people can also be treated in a similar manner where individuals with certain characteristics exist in a virtual world and interact. These types of models where every individual is modeled independently are known as an agent-based models. By treating individual people as stochastic agents that follow a set of rules and probability distributions we can understand large groups of people and their behaviour as a community. The properties of the community and its demographics are determined by the rules which govern single agents. In the same way that many collision events between many particles might yield a distribution describing the mass distribution of say, the Higgs boson, a simulation of many people over several days can yield distributions providing

insight into attendance rates or social mixing rates at various locations in a virtual world. In an event generator we gain insight into various physical processes, potential new physics or rare decays where-as in an agent-based model we analogously gain an understanding of how different demographics of people interact in different locations in the simulated virtual world. Therefore, we can employ a Monte Carlo simulation to understand social mixing of different demographics.

In this thesis we apply these two types of Monte Carlo simulations in two different domains which we will explore in the two sections of this chapter. In Chapter 5 we will use SHERPA a general purpose event generator as a Standard Model simulator with which we can constrain our understanding of the Yukawa Charm coupling. Then in Chapter 7, we use and build upon JUNE [23] an agent based Monte Carlo epidemiological framework for modelling social interactions.

3.1 Monte Carlo Event Generators for Particle Physics

A Monte Carlo event generator is a tool in which we aim to simulate collision events that occur at various particle colliders using stochastic methods. They are particularly useful as they provide a way to test our models against experimental observations, develop new models, create new more robust analysis tools and allow a greater flexibility and control over processes we want to study. There exist a large selection of generators, HERWIG [24, 25], PYTHIA [26] and SHERPA [27] to focus particularly on the high energy physics general purpose event generators. The chronology of event generation can be broken down into the following steps:

1. Hard process: The set of matrix elements are generated (e.g. COMIX [28], AMEGIC++ [29]) that define a process are used to sample phase-space points – a set of incoming and outgoing particles and their four momentum – which are possible under conservation laws and the relevant parton distribution functions

(e.g. NNPDF 3.0 PDFs [30] from LHAPDF [31]). These phase space points are sampled in a clever way with algorithms such as RAMBO [32] which gives each phase-space point a probability based on the particle masses and their distribution in space. This probability provides a weighting which relates the phase-space point to its contribution to the total cross section of the process. These weightings can be used to systematically reject or accept space-space points such that the full phase space is probed efficiently, we call an accepted point an event and the set of rejected and accepted events trails.

2. Parton shower: The parton shower describes the evolution of the high energy hard scale down to the hadronisation scale (e.g. CSSHOWER [33]). This is the regime where perturbation theory and our understanding of QED and QCD can be implemented. We model the parton shower with an evolving scale, the choice varies by event generator some choices include virtuality, perpendicular momentum or angular distance. This scale is iteratively reduced by allowing the partons to undergo splitting in which they produce softer and more collinear radiation until all of the particles are below the cut off hadronisation scale $\sim$1 GeV. The quarks shower gluons but the gluons can themselves shower gluons or split into quarks generating new flavour content. The splitting process is defined by the Dokshitzer–Gribov–Lipatov–Altarelli–Parisi (DGLAP) equations and the Altarelli-Parisi splitting functions [34–36]. This technology also models the initial state radiation by backwards evolving the input partons of the hard process up to the original scale of the partons in the beam.

3. Hadronisation: When we reach the hadronisation scale, perturbation theory breaks down and thus we have to consider empirical non-perturbation models such as the string (Lund string model [26]) or cluster models [37]. The final products are required to be colourless due to confinement therefore the partons have to be clustered into colour neutral objects such as baryons and mesons.

4. Hadronic decays: The hadrons produced may not have a long lifetime and they

may decay within the detector, it is typical that the majority of the detected decay products end up as pions and kaons (e.g. SHERPA's built-in models for hadron decays).

5. Underlying event: Aside from the hard process we also have to consider the remnants of the hard process which can themselves interact and therefore undergo all or some of the above – which should not be overlooked. This can be incorporated in multiple parton interaction models (e.g. AMISIC++ [38]).

6. Detector simulation (optional): Lastly, the final four momentum objects produced in the event generator are not representative of real data for several reasons:

   - Detector geometry: Detector components do not fully enclose the interaction vertex as certain components do not span all values of η.
   - Spacial-resolution: The calorimeter cells have a finite size thus, we have a minimum resolution in (η, ϕ) for which we can place particles or distinguish them.
   - Instrumentation effects: The measurements of induced currents from particle interactions with the detector are subject to digitisation error and there is inherent electrical noise which leads to an uncertainty in the energy measurements.

   There exist tools such as DELPHES [39] which can estimate these effects with energy smearing and position discretisation by summing the energy of tracks in the discrete calorimeter cells.

We can see these steps outlined in Fig. 3.1

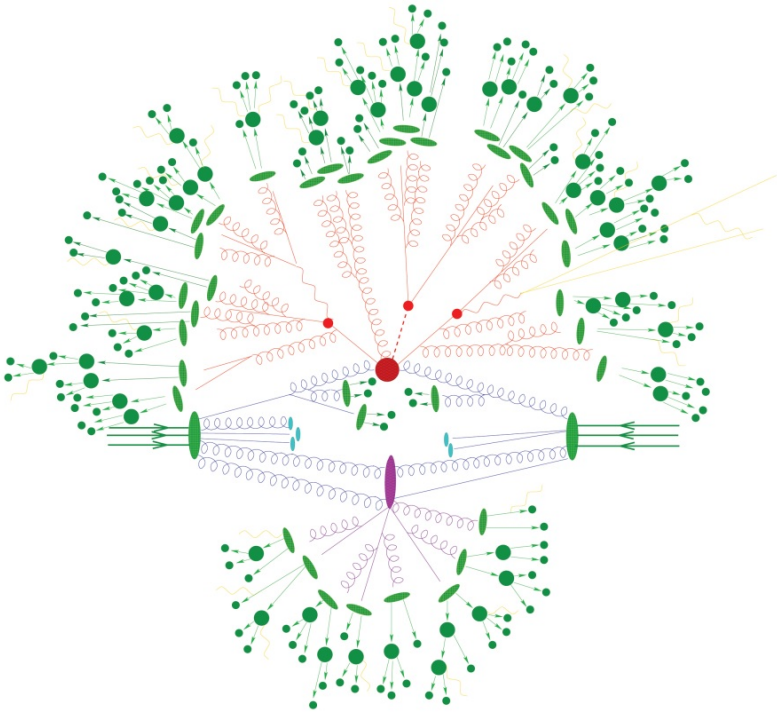


Figure 3.1: Schematic of a hadron-hadron collision at a hadron collider such as the LHC. The incoming hadrons are indicated by the three green arrows representing the valence quark content of the beams. The red blob at the centre indicates the hard process. The initial state and final state radiation are shown in blue and red respectively which are directly connected to the hard process which constitutes the parton shower. Hadronisation is shown by the green blobs after the final state radiation and then their subsequent decays from further green blobs. We see the underlying event indicated by the purple and cyan blobs. Figure from [5].

3.2 Monte Carlo Simulations for Social Mixing

Agent based models are a class of simulations in which we model the behaviour of some system by introducing a set of rules and stochastic conditions which determine the behaviour of individual interacting agents. These types of models are particularly powerful for understanding disease spread and social interactions and developing epidemic mitigation strategies [40, 41]. By simulating any scenario from the bottom-

up, we gain an understanding of the emergence of macroscopic properties from microscopic ones belonging to the agents. In particular, we will look at agent based models in which the agents are individual people which interact within a virtual world in various virtual settings. We make use of the JUNE framework [23] which was built initially to model the COVID-19 pandemic in the United Kingdom as an independent open-source epidemic model. We will extend the JUNE framework with a social interaction tracking class which takes virtual contact surveys at every simulated time step at every virtual venue. JUNE was build with flexibility in mind such that it can be extended to further settings [42]. JUNE has a large number of parameters that can be individually changed. Census information, social behaviour and enforced pandemic measures can all be customised for unique locations or social behaviours. Many hyper parameters can be tuned to the emerging macroscopic epidemiological behaviours such as case, death and hospitalisation rates to build a reliable epidemiological simulation. Alternatively various social policies such as lock downs or social distancing can be studied for better pandemic mitigation strategies.

Here we present a general outline of the virtual world construction in agent based models. More details can be found in Section 7.2.2 with more specifics of JUNE.

1. Geography and demography: The virtual population is created using geographical census data. The framework divides the geography into a hierarchy of three layers:

   - Regions: Areas on the scale of counties.
   - Super Areas: Areas containing a few thousand households.
   - Areas: Areas containing a few hundred households at most.

   The virtual population is constructed by creating the individuals represented in the census in which each agent has a sex, age and a residential area. Further attributes can often be assigned if required (and available) such as sexuality, gender and race. In the location of interest in this project, the only available data are those of sex, age and residential area.

2. Household construction: The first venue that the agent will attend is that of their household. The virtual population is sorted into households of varied types depending on the distribution of typical household properties and types. These could include internal household properties such as:

   - people per household,
   - number of children,
   - spousal age gap,
   - mother child age gap,

   or external properties:

   - number of nuclear households,
   - number of single parent households,
   - number of multi-generational households.

   In Section 7.2 we will detail this procedure.

3. Venue construction: The agents can move around the simulation between a selection of social settings in which they can interact with other agents outside of their household. There are two classes of venues:

   - Assigned venues: Venues which only specific agents are required to attend on a regular basis, these include workplace or educational settings.
   - Unassigned venues: Venues that any agent can attend with a probabilistic chance depending on their individual characteristics. Examples include shops and recreational spaces like cinemas, pubs or community centres.

   All of the venues are assigned to geographical areas in the model such that agents generally travel to venues local to them. Assigned venues have specific agents allocated to that location and they must attend (unless overridden by social policies or rules e.g. sickness) at the allotted times.

After the virtual world is constructed, simulations can commence. The simulation time is divided into discrete timesteps within calendar days. The calendar date system allows for behavioural differences between weekdays and weekends, this is particularly important for the workplace and educational settings. An individual in a typical day is expected to attend an assigned venue, such as 8 hours at work before they then attend in other venues and return home for the night and remaining hours. While the assigned venue is predetermined, the other activities are chosen based on a stochastic process. The probabilities for attending a venue type are modelled on a Poisson process which depends on the agents characteristic properties. The particular venue is chosen randomly from the closest of that type of venue up to some travel limit.

At each time step, the social interaction of the agents at each location in the virtual world can be modelled. Social mixing can be quantified into social mixing matrices Δ_{ij} which is the number of contacts in a characteristic time between person of characteristics i that contacts person of characteristics j. Contact matrices that represent social mixing patterns in populations are a vital input to epidemiological models [43,44]. There are many choices of the type and format of these social mixing matrices, we will address several in Chapter 7 they can depend on normalisation, characteristic binning and social mixing type [45–47]. Traditionally, contact matrices are derived using large scale surveys, where participants record the number of contacts they have in different locations and the ages of the people they came into contact with [48]. In Chapter 7 we will develop a novel mixed-method approach which combines an agent based model with a light-weight contact survey.

Naturally, agent based models come with a large selection of hyper-parameters which define the virtual world and the agents behaviour. In this project we ignore the epidemiological elements of JUNE and the parameters associated to the disease spread specifically. We focus on tuning parameters relating to that of venue demographics agent behaviour such that they are representative of reality by comparing against survey findings conducted in our particular setting.

Chapter 4

Particle Physics

In Chapter 5, we will concern ourselves with methods to improve current constraints on the upper-bound of the charm Yukawa coupling y_c and provide projections on expected confidence limits as we move into the high luminosity phase of the LHC. In this chapter we will present the required requisite physics for this goal. A presentation of the Standard Model (SM) of particle physics can be found in the Appendix A.

4.1 The Charm Yukawa Coupling

In the SM without the Higgs boson there exists no way to introduce mass terms for the charm quark in a gauge invariant way. In order to generate mass for these particles we must introduce a scalar field with the following distinctive Lagrangian term – the Higgs sector Lagrangian,

$$\begin{aligned} \mathcal{L}_{\text{Higgs}} &= (D_\mu H)^\dagger (D^\mu H) - V(H) \\ &= (D_\mu H)^\dagger (D^\mu H) - \mu^2 H^\dagger H - \lambda (H^\dagger H)^2 \ , \end{aligned} \tag{4.1.1}$$

where we have the covariant derivative,

$$D_\mu = \partial_\mu - i g_Y \frac{Y}{2} B_\mu - i g \frac{\sigma^i}{2} W_\mu^i \ . \tag{4.1.2}$$

The Higgs field has a scalar potential that depends on two parameters, μ and λ such that if $\mu^2 < 0$ then there exists a minima in the potential at a non-zero value of the field. The scalar field transforms in a non-trivial way under $\mathrm{SU(2)_L}$ as a doublet made up of two scalar complex fields, ϕ^+ and ϕ^0 in terms of four real scalar fields, ϕ^1, ϕ^2, ϕ^3 and ϕ^4;

$$H = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix} . \tag{4.1.3}$$

Using Eq. (4.1.1) we can show that the potential is minimised for $\mu^2 < 0$ when,

$$|H|^2 = H^\dagger H = -\frac{\mu^2}{2\lambda} = \frac{\nu^2}{2} . \tag{4.1.4}$$

Where we have defined a new quantity ν, the vacuum expectation value (VEV) of the Higgs boson potential and it is experimentally determined to be $\nu \approx 246\,\mathrm{GeV}$ [49]. The Higgs field can be re-written under the unitary gauge where ϕ^3 is chosen to align in the radial direction of the potential and thus h are perturbations uphill from the VEV;

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h \end{pmatrix} . \tag{4.1.5}$$

There is actually flexibility in the gauge choice, the four real fields imply there are in fact four degrees of freedom, three of these correspond to modes which move along the equipotential of the Higgs potential which give rise to three massless Goldstone boson modes. In gauging away these modes the missing degrees of freedom act as the longitudinal modes of the SM gauge bosons giving them mass. The final mode (moving in the radial direction) feels a quadratic potential where perturbations, h are understood to be excitations corresponding to a massive particle, the Higgs boson. We can show by substitution into Eq. (4.1.1) that coefficients in the quadratic terms of the Higgs boson, h^2 equate to the Higgs boson mass,

$$m_h = \sqrt{2\lambda\nu^2} . \tag{4.1.6}$$

We can further introduce an interaction term between the Higgs boson and the

fermions – the Yukawa sector Lagrangian.

$$\mathcal{L}_{\text{Yukawa}} = -(y_u)_{ij}\bar{q}^i_{\text{L}}\tilde{H}u^j_{\text{R}} - (y_d)_{ij}\bar{q}^i_{\text{L}}Hd^j_{\text{R}} - (y_e)_{ij}\bar{l}^i_{\text{L}}He^j_{\text{R}} + \text{h.c.}\,, \tag{4.1.7}$$

where the indices i and j run over the generations of particles. In Eq. (4.1.7) we introduce the conjugate Higgs doublet, $\tilde{H} = i\sigma^2 H^*$ to extract the up type quark fields and we use the Higgs double H to extract the down type quark fields and leptons. Looking explicitly at the first generation up quark and by substituting in the Higgs doublet and its conjugate,

$$\begin{aligned}\mathcal{L}_{\text{Yukawa}} &= -(y_u)_{11}\bar{q}^2_{\text{L}}\tilde{H}u_{\text{R}} - (y_u)_{11}\bar{u}_{\text{R}}\tilde{H}^\dagger q^2_{\text{L}} \\ &= -\frac{(y_u)_{11}(h+\nu)}{\sqrt{2}}(\bar{u}_{\text{L}}u_{\text{R}} + \bar{u}_{\text{R}}u_{\text{L}}) \\ &= -\frac{(y_u)_{11}h\bar{u}u}{\sqrt{2}} - \frac{(y_u)_{11}\nu\bar{u}u}{\sqrt{2}}\,,\end{aligned} \tag{4.1.8}$$

it is possible to read off the mass term, $m_u = y_u\nu/\sqrt{2}$, where y_u is the Yukawa coupling for the up quark. It is straightforward to repeat this process for the remaining fermions and retrieve similar expressions. Alternatively, we can consider the Yukawa coupling as a function of the mass,

$$y_i = \frac{\sqrt{2}}{\nu}m_i\,. \tag{4.1.9}$$

As exemplified in Eq. (4.1.9) the Yukawa coupling is directly proportional to the mass of the particle concerned. Given the hierarchy of the fermion masses this poses a significant challenge in the determination of the Yukawa coupling to the second-generation quarks which are considerably lighter than much of the massive content of the SM. In addition, the event topologies and in particular the jet structures are not particularly distinctive for processes involving $H_{\to c\bar{c}}$ decays. It is particularity difficult to extract a significant number of $H_{\to c\bar{c}}$ events as compared to processes involving $H_{\to b\bar{b}}$ due to weak event features and the relatively small branching fraction. Background processes with cross sections many orders of magnitude larger than any $H_{\to c\bar{c}}$ process can span similar areas of phase-space in significant numbers therefore making them difficult to eliminate in a cut-flow analysis.

4.2 Collider Physics

The LHC is the latest particle collider build at CERN it is a 27km ring of super conducting magnets which accelerates protons to ultra relativistic speeds. Bunches of these protons collide at key locations around the storage ring where detectors are placed. Two such detectors are the "A large ToroidaL ApparatuS" (ATLAS) [50] and "Compact Muon Solenoid" (CMS) [51] detectors. These detectors have an array of instruments which determine the energy, mass and charge of the outgoing particles from the collisions.

The collision of two bunches of protons produces extremely complex structures which can be understood using our understanding of the SM or new physics models. This defines particle physics phenomenology – the application of theoretical physics to understand experimental data – one such success story is the discovery of the Higgs boson. The inclusion of the Higgs boson in the SM was very strongly theoretically motivated and it was hoped that Tevatron, LEP or (more likely) the future LHC would discover this particle – which it did in 2012 [52, 53]. Further, phenomenology can be utilised to build Monte Carlo models (Section 3.1) which simulate the collisions at real-world or future colliders thus contributing to further tests of the SM or beyond the Standard Model (BSM) models.

There are a few key points that need to be discussed to understand precisely how colliders differ from one another. Firstly, consider the beam constituents, if only leptons are used then any processes relating to QCD and the strong force are much less accessible than in hadron colliders. However, the beam choice of composite particles like protons comes at the expense of an uncertainty of the flavour and momentum of the particles interacting in the hard process. The collision of a parton from within the proton carries only a fraction of the total energy of the proton which can be modelled with parton distribution functions (PDFs). The centre of mass energy determines kinematic viability of processes and it is defined as the sum of the energies of the colliding particles in the centre of mass frame (e.g. protons at

the LHC),

$$E_{\mathrm{CM}} = \sqrt{s} = \sqrt{(p_1 + p_2)^2} \, . \tag{4.2.1}$$

Here we introduce s, the Mandelstam invariant which is the sum of the two beam four-momenta, p_1 and p_2 squared. The four momentum, $p = (E, \boldsymbol{p})$ is defined by the beam particle energy, E and three momentum, $\boldsymbol{p}$. Again, in the case of a proton proton collision, where the partons involved in the hard-process collision take a fraction of the energy of the beam, the actual centre of mass energy of the collision is much less and is also uncertain. The LHC is described as having a centre of mass energy of 14 TeV however individual collision events only have a centre of mass energy of a fraction of this. Further, it is not known which constituents have participated in the collision, this has to be inferred from the event structure and particles detected in the final state. Alternatively, a collider involving leptons which have no internal structure and which only interact via the electromagnetic and weak forces therefore provide a much "cleaner" collision. However, in a circular collider it is important to discuss synchrotron radiation; charged particles of mass, m_0 and energy, E moving in a circular path of radius, R will radiate energy at a rate per turn,

$$\delta E \propto \frac{E^4}{m_0^4 R} \, . \tag{4.2.2}$$

Therefore building a circular collider of equivalent energy for a electron collider compared with proton collider is significantly harder as the rate of loss of beam energy is astronomically higher.

Luminosity is a measure of the instantaneous rate of collision events per unit time per cross section. The total number of events for a given instantaneous luminosity $\mathcal{L}$ is,

$$N = \sigma(s) \int \mathcal{L} dt \, . \tag{4.2.3}$$

The cross section, $\sigma(s)$ is the probability an event will occur at a given centre of mass energy s. It is more common to discuss the integrated luminosity which is a measure of the accumulated data at a collider and as it is independent of the cross section. It

is simply a metric of the total amount of data delivered. The integrated luminosity naturally increases over run time, but also increases with increased instantaneous luminosity. Run 1 of the LHC concluded with an integrated luminosity delivered of $20\,\mathrm{fb}^{-1}$, Run 2 with $150\,\mathrm{fb}^{-1}$ and Run 3 is expected to reach $300\,\mathrm{fb}^{-1}$. By 2028 it is expected that the LHC will move into the High-Luminosity phase of the LHC in which it is expected to achieve an integrated luminosity of $3000\,\mathrm{fb}^{-1}$. The shear scale of the statistics that this will provide will yield new opportunities to scrutinise processes under greater sensitivity [54].

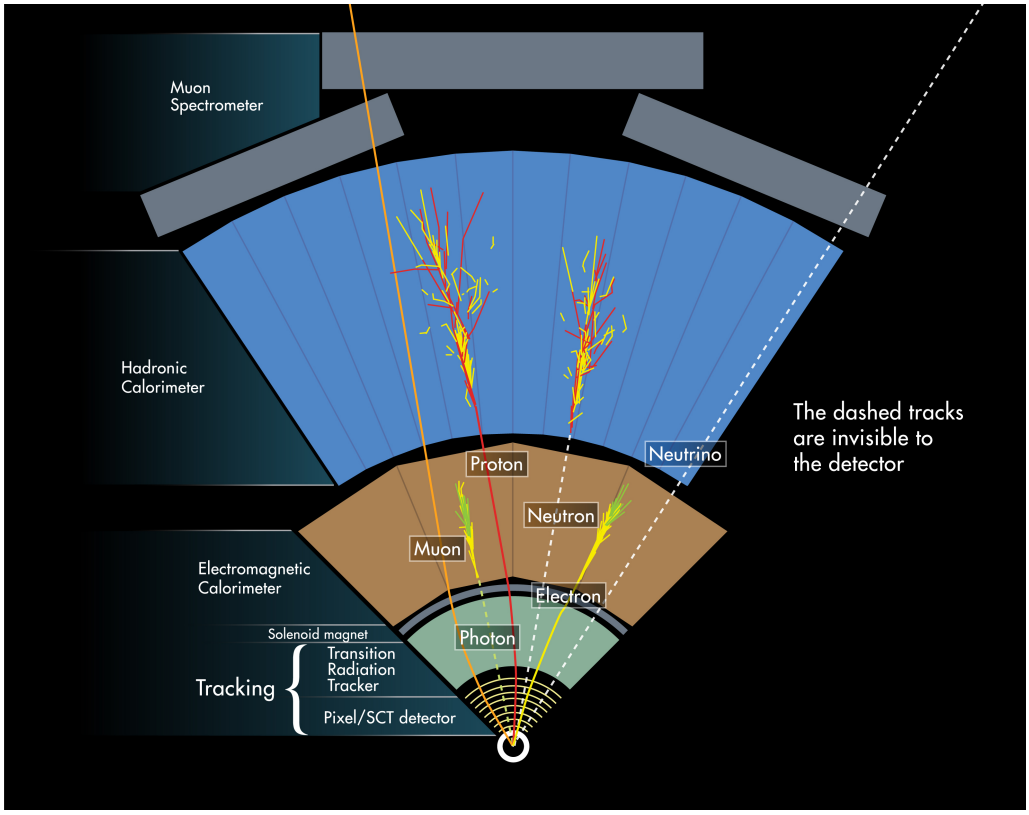


Figure 4.1: A cross-section schematic of the ATLAS detector at the LHC [55]

Collision event data is measured at key locations around the storage rings – the detectors. A cross sectional schematic of the ATLAS detector is shown in Fig. 4.1 where sections of the detector are clearly distinguished into layers [56]:

- Tracking Chamber (Inner Detector),
- Electromagnetic Calorimeter (ECAL),

- Hadronic Calorimeter (HCAL),

- Muon Spectrometer.

The tracking chamber measures the charged particles, their direction, momentum and magnitude of charge. This part of the detector is also used to determine impact parameters of particles to a vertex (see Section 4.3) and determine the primary vertex of an event, this is particularly useful as it can be used to mitigate and segment pile-up information. Many bunches of protons collide per second which can cause multiple collision events per bunch crossing. The ability to distinguish individual collision events via primary vertex fitting is critical. ECALs are designed to absorb most of the particles from the collision, those which interact with scintillation material and produce photons which can be measured. This portion of the detector is most sensitive to electrons, positrons and photons. HCALs are designed to measure the hadrons which pass through the ECAL unperturbed. These hadrons collide with the nuclei in a dense absorber material which can produce secondary particles. These particles interact with scintillation material producing photons which can be detected. The decay of these particles is described as a shower which forms structures such as jets (see Section 4.4) which can be used to examine the identity of the original hadron producing them. Lastly, we have the Muon Spectrometer which measures precisely the energy and location of muons which thus far have not interacted with the detector. This portion of the detector is the largest and outermost portion. All the information collected by the detector can be used to reconstruct a picture of an event.

The detector is designed as a cylindrical barrel and therefore we map the direction of particle tracks to a cylindrical co-ordinate system. We define the beam axis to be along the z-axis such that event data is transitionally invariant and we define two angles, the azimuthal angle ϕ and the polar angle, θ. A natural set of co-ordinates to consider that can be mapped onto the surface of the detector are ϕ and the

pseudorapidity η,

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) . \tag{4.2.4}$$

In particular, differences in these quantities ($\Delta\phi$ and $\Delta\eta$) are Lorentz invariant for boosts along the beam axis for relativistic particles. The rapidity difference Δy is always Lorentz invariant for boosts along the beam axis for any particle where the rapidity y,

$$y = \frac{1}{2}\ln\frac{E + p_z}{E - p_z} . \tag{4.2.5}$$

We choose to use $\Delta\eta$ such that we can define a measure of angular separation, ΔR which is a Lorentz invariant as a function of angular quantities only. This quantity will be useful for defining jets,

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} . \tag{4.2.6}$$

Particles which have large high transverse momentum compared with their rest mass can be considered relativistic. The transverse momentum is given by,

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} . \tag{4.2.7}$$

The two incoming beams have equal and opposite momenta entirely in the z axis thus we know that the total momentum in the frame of the detector is zero. We therefore know if $\sum p_{\mathrm{T}} \neq 0$ then we have particles that the detector has failed to detect, which could be neutrinos or alternatively BSM particles, we call this difference from zero the missing momentum. These quantities can be used to define the dimensions of the detectors such that we can build realistic and representative analysis in Monte Carlo generators. For example the Inner Detector of ATLAS is unable to perform vertex determination with tracks with $p_{\mathrm{T}} < 0.1\,\mathrm{GeV}$ and $|\eta| > 2.5$ due to the physical dimensions of the detector barrel and instrumentation limitations. The Monte Carlo event generators provide a further avenue for validation of theories of physics we wish to test and we will discuss them further in Chapter 3.1.

4.3 Vertex Fitter

In order to fit useful observables such as invariant masses we need to understand which particles belong to a particular event otherwise the data is contaminated by pile-up effects. An interesting high-energy collision could be contaminated by other soft collisions or from other collision events between different protons a single proton-proton bunch crossings, which makes it difficult to reconstruct the objects in each event final state. One of many techniques used to tidy the collision data is vertex fitting, particles sharing the same primary vertex are considered to be constituents of a single particle collision event. In the case of Monte Carlo simulations the primary vertex of an event is known by definition, the origin of the co-ordinate system $(0, 0, 0)$. However, in the case of the LHC this position needs to be determined so that the separate events in pile-up with different primary vertices can be distinguished. The current method of vertex fitting is adapted from the Adaptive Multi-Vertex fitter (AMVF) algorithm [57] which takes the reconstructed tracks and an estimate vertex position and iteratively improves upon the vertex position and gradually down weights potential outliers (tracks that do not share that vertex). In Fig. 4.2 we see a schematic for a single event which is defined by its primary vertex which all the particles in that event share. The figure also shows secondary vertices which result from the decay of short lived particles produced in the primary vertex. Higher order vertices can also be produced where particles produced from decays may themselves decay.

In order to fit a vertex from a set of tracks we have to define a loss function which assesses the quality of fit. We define the least squares estimator which sums the (standardised) impact parameters $d_i(\boldsymbol{v})$ of all n tracks from the vertex position $\boldsymbol{v}$,

$$L(\boldsymbol{v}) = \frac{1}{2}\sum_{i=1}^{n} \chi_i(\boldsymbol{v})^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{d_i(\boldsymbol{v})^2}{\sigma_i^2} \ . \tag{4.3.1}$$

The impact parameters are the perpendicular distance between the point of closest approach of the particle track and the vertex and σ_i is an uncertainty of the impact

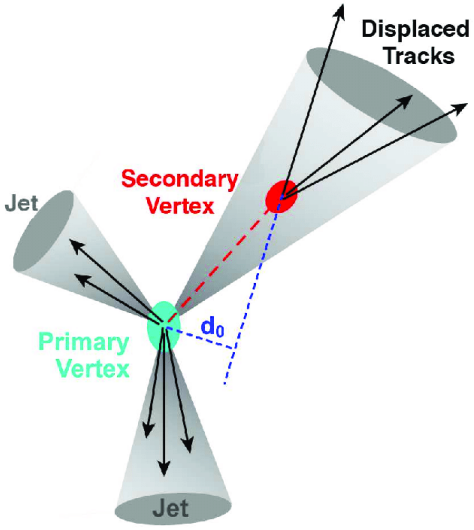


Figure 4.2: Schematic of a a primary and secondary vertex. The green blob represents the hard process and primary vertex. The red blob a secondary displaced vertex. d_0 indicates the impact parameter of a particular final state particle to the primary vertex, indicating its origin is from a displaced vertex. Figure from [58].

parameter. This loss function can be minimised by taking the derivative with respect to $\boldsymbol{v}$ and setting it to zero,

$$\frac{\partial L(\boldsymbol{v})}{\partial \boldsymbol{v}} = \sum_{i=1}^{n} \chi_i(\boldsymbol{v}) \frac{\chi_i(\boldsymbol{v})}{\partial \boldsymbol{v}} = 0 \, . \tag{4.3.2}$$

This expression can be solved for the optimal vertex position exactly or using numerical techniques such as gradient descent or more sophisticated methods like Kalman filters [59]. This fitting algorithm can be improved upon by weighting the tracks by down-weighting tracks that are less likely to be associated to the vertex instead of outright rejecting them. The weights are defined in terms of χ_i for each track i,

$$\omega_i(\chi_i^2) = \frac{e^{-\chi_i^2/2T}}{e^{-\chi_i^2/2T} + e^{-\chi_C^2/2T}} \, . \tag{4.3.3}$$

Eq. (4.3.3) depends on two hyper parameters, a temperature T which controls the shape of the weight distribution and a cut off χ_C which defines the threshold value where the weight should equal 0.5 i.e. where a track is considered more likely to be

an outlier than not.

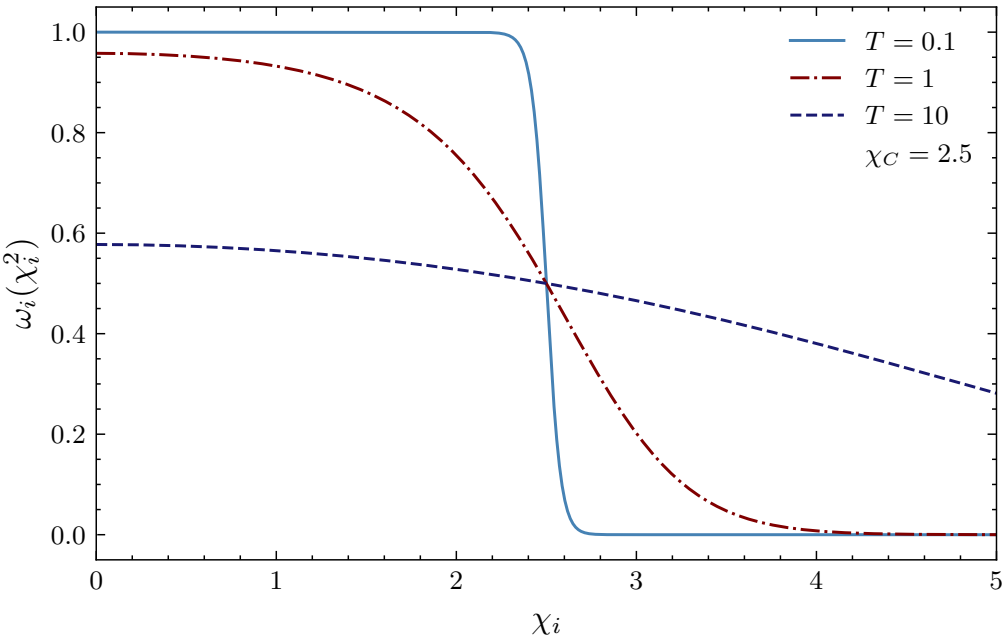


Figure 4.3: The distribution of weights for a standardised distance, χ_i for a selection of temperatures, T which all share the same cut off threshold of $\chi_C = 2.5$.

T is reduced during the procedure as an annealing scheme which helps the vertex fitter fall into a global minimum. Eq. (4.3.2) can now be modified to incorporate the weights,

$$\frac{\partial L(\boldsymbol{v})}{\partial \boldsymbol{v}} = \sum_{i=1}^{n} \omega_i(\chi_i(\boldsymbol{v})^2)\chi_i(\boldsymbol{v})\frac{\chi_i(\boldsymbol{v})}{\partial \boldsymbol{v}} = 0 \ . \tag{4.3.4}$$

This equation can not be solved exactly so we move to iterative procedures in which an initial vertex position provides a set of weightings which then provides a new improved vertex position. The procedure is repeated until a convergence stops or a desired precision is reached. This vertex fitting methodology can be use to fit primary, secondary and even tertiary vertices for a single event (cf. Fig. 4.2), we apply the same procedure for higher order vertices by considering a subset of the particle tracks – for example a jet suspected to be from a b-hadron decay – then we can retrieve these higher order vertices.

4.4 Jet Clustering

As unstable and highly energetic particles propagate they can decay into a plethora of hadrons or soft partons and photons. This collection of particles tend to be clustered together within η, ϕ space and we call this collection of particles a jet. To define a jet quantitatively we need to group together the particle tracks in a way that is not dependent on the soft or collinear emissions i.e. it is infrared and collinear safe (see Section 4.5). This is particularly important for Monte Carlo simulations in which we do not want the jet algorithms to be dependent on stochastic processes of the parton shower as that could lead to ambiguity in the number of jets or their properties for example the addition of a soft gluon. Jet algorithms take particles within a radius R and iteratively combine pairs of them into quasi-particles until certain criteria are reached; the remaining quasi-particles define the jets. There exist a selection of jet clustering algorithms and here we will focus on the general class of k_{T} sequential algorithms. These cluster particles recursively based on its transverse momenta. We define two variables in momentum space,

$$d_{iB} = p_{\mathrm{T},i}^{2p} \, , \tag{4.4.1}$$

for a distance of one particle to the beam i and

$$d_{ij} = \min\{d_{iB}, d_{jB}\}\frac{\Delta R_{ij}}{R} \, , \tag{4.4.2}$$

for a pair of particles (or beam) i, j. The clustering proceeds iteratively where any objects with the smallest d_{ij} are clustered, if $d_{iB} < d_{ij}$ then i is clustered with the beam otherwise i, j are clustered together. This is repeated until the smallest d_{ij} exceeds some cut off value, d_c which defines the minimal separation between each jet to one another and to the beam. This algorithm is parameterised by the cut off d_c, the cone-like size R and lastly the exponent of the momentum p. The exponent parameterises different k_{T} type algorithms:

- $p = 1$: The k_{T} algorithm [60],

- $p = 0$: The Cambridge-Aachen algorithm [61],
- $p = -1$: The anti-k_{T} algorithm [62].

Each of these paradigms of jet clustering provides a unique clustering of tracks into jets. The k_{T} algorithm is the original sequential jet clustering algorithm and it clusters particles preferentially in order of increasing transverse momentum. The Cambridge-Aachen algorithm clusters by proximity in η, ϕ space and it is not a dependent to the magnitude of the momentum. These two methods cluster particles into relatively irregular shapes, whereas the anti-k_{T} algorithm provides more regular shapes by clustering from hardest particle to the softest. These sequential type jet clustering algorithms are infrared and collinear safe and in particular the anti-k_{T} algorithm produces regular conical jet shapes. Further, the extracted jet-axis from the anti–k_{T} algorithm is insensitive to soft radiation therefore it simplifies the calorimeter energy calibration procedure. These properties mean it has been adopted as the default choice at the LHC. The choice of jet size R and d_c varies depending on the experiment or analysis constructed, for example when looking for Higgs fat jets, we look for relatively large jets (i.e. $R \approx 1$) and require that the transverse momenta of the jet exceed the rest mass (i.e. $d_c > 125\,\mathrm{GeV}$). After clustering the events into jets we can define further observables in the event based on jet structure and displaced vertices from short lifetime hadrons decays. These are useful to conduct jet flavour tagging.

4.5 Observables

There is a plethora of potential observables that are used to glean insight into the underlying hard process of an event. In order for an observable to be insightful we look for observables that are infrared and collinear safe, this is particularly important for Monte Carlo simulations where we do not want our results to be dependent on stochastic effects.

- Infrared Safe: Features are not dependent on the introduction of soft (low energy) radiation. For example the addition of a low energy gluon in the parton shower does not appreciably change a feature in the final state.

- Collinear Safe: Features are not dependent on the introduction of a collinear (or nearly collinear) particle on top of the path of another particle. For example collinear splitting of a particle into two particles.

It is important that these requirements are met as it allows for easier comparison of Monte Carlo data to experiment data. In the work in Chapter 5 there are a large number of observables that are investigated, although not all that we will define are ultimately used we will define all of them for completeness. The observables can be separated into a few categories, global observables, jet observables and vertex observables. We consider such a expansive selection of observables so that we can exploit many subtle differences between the signal processes and the backgrounds we which to eliminate. Not all of the features that are considered are infrared or collinear safe.

4.5.1 Global Observables

Observables which characterise the whole event and its topology.

- Particle multiplicities: N_{P}. Number of reconstructed particle tracks, can be all particles or subsets of charged ($N_{\mathrm{P}^{\pm}}$) or flavoured ($N_{\mathrm{P}^{f}}$). Which is not a collinear and infrared safe observable. This observable is not used in our cut-flow.

- Number of jets: N_{jets}. Total number of jets in an event defined by R, d_c and p.

- Missing transverse momentum: $E_{\mathrm{T,\,miss}}$. The total energy undetected from invisibles such as neutrinos.

- Number of isolated leptons: $N_{\text{iso }\ell}$. Total number of isolated leptons which are defined by the following criteria. For each identified lepton track we demand that the total transverse energy of all particles in a cone of size R_{iso} around the lepton direction is constrained to not exceed 5% of the lepton transverse energy $E_{T,\ell}$,

$$\sum_{\Delta R_i < R_{\text{iso}}} E_{T,i} \;\leq\; 0.05 \cdot E_{T,\ell} \;. \tag{4.5.1}$$

  This feature is infrared and collinear safe as the isolated leptons can be constrained to be sufficiently hard that the associated reconstructed leptons are directly mapped to the hard process.

- Number of displaced vertices: N_{vertex}. The total number of reconstructed displaced vertices defined by the T annealing scheme and χ_C.

4.5.2 Jet Observables

Observables which characterise an individual jet.

- Particle multiplicities: $N_{\text{P}\in\text{jet}}$. Number of reconstructed particle tracks contained within the jet, can be all particles or subsets of charged ($N_{\text{P}^{\pm}\in\text{jet}}$) or flavoured ($N_{\text{P}^{f}\in\text{jet}}$). Which is not a collinear and infrared safe observable. This observable is not used in our cut-flow.

- Tagged: f_{tag}. Is the jet tagged to contain t, b, c or μ. If the jet can be associated to a hard process this feature can be infrared and collinear safe.

- Jet clustering distance ratios: $D^{n\to n-1}$. During each iterative step in the jet clustering algorithm we calculate distances, d_{ij} for the clustering n to $n-1$ quasi-particles. We can re-cluster and record these distances and normalise them with respect to the transverse momentum of the original jet to the beam, $p_{\text{T, jet}}$,

$$D^{n\to n-1} = \frac{\sqrt{d_{ij}^{n\to n-1}}}{p_{\text{T, jet}}} \;. \tag{4.5.2}$$

In particular we consider;

  - $D^{3\to2}$,
  - $D^{2\to1}$,
  - $D^{3\to2}/D^{2\to1}$.

- Momentum: p^{μ}_{jet}. Four momentum of the reconstructed jet. The four momenta can be used to define the following further features:
  - Energy: $E_{\text{jet}} = E_{\text{P}\in\text{jet}} = p^{0}_{\text{jet}}$ or energy contained in charged particles $E_{\text{P}^{\pm}\in\text{jet}}$,
  - Transverse jet momentum: $p_{\text{T, jet}}$,
  - Invariant Mass: $m_{\text{jet}} = \sqrt{p^{2}_{\text{jet}}}$,
  - Pseudo-rapidity: η_{jet},
  - Azimuthal angle: ϕ_{jet}.

- Transverse momentum in jet: $p_{\perp\in\text{jet}}$. The sum of the transverse momentum of the track constituents of the jet with respect to the unit vector axis of the jet, $\hat{\boldsymbol{p}}_{\text{jet}}$.

$$p_{\perp\in\text{jet}} = \sqrt{p_x^2 + p_y^2} \quad \ni \quad \hat{\boldsymbol{p}}_{\text{jet}} \cdot \hat{\boldsymbol{z}} = 1 \; . \tag{4.5.3}$$

- Pull: $\boldsymbol{t}$ [63]. A vector which describes the connectedness of subjets within a larger jet. Its magnitude $|\boldsymbol{t}|$ or its parallel ($\boldsymbol{t}_{\parallel}$) and perpendicular ($\boldsymbol{t}_{\perp}$) projection to some axis can be calculated.

$$\boldsymbol{t} = \sum_{i\in\text{jet}} \frac{p_{\text{T},i}(\Delta R'_i)}{p_{\text{T, jet}}} (\Delta \boldsymbol{R}'_i) \; , \tag{4.5.4}$$

where $(\Delta \boldsymbol{R}'_i) = (\Delta y_i, \Delta\phi_i)$ with the Δx the difference between the constituent i and jet values of x and $\Delta R'_i$ its magnitude which is a new angular separation with y the rapidity.

- Angularities: λ_a^x [64]. These generalised angularities probe the angular, energetic structure of jets,

$$\lambda_\beta^x = \frac{1}{(p_{\mathrm{T,\ jet}})^x R^\beta} \sum_{i \in \mathrm{jet}} (p_{\mathrm{T},i})^x \Delta R_i^\beta \ . \tag{4.5.5}$$

- Eccentricity: ϵ [65]. Describes the deviation of the energy jet profile from a perfect circle. We define a matrix M,

$$M = \sum_{i \in \mathrm{jet}} \begin{pmatrix} (\Delta\eta_i)^2 & (\Delta\eta_i)(\Delta\phi_i) \\ (\Delta\phi_i)(\Delta\eta_i) & (\Delta\phi_i)^2 \end{pmatrix} \ . \tag{4.5.6}$$

  Eccentricity is defined by the two eigenvalues of M which we order by their magnitude, $\lambda_{\max} > \lambda_{\min}$,

$$\epsilon = 1 - \frac{\lambda_{\min}}{\lambda_{\max}} \ . \tag{4.5.7}$$

  A value of $\epsilon = 0$ describes the energy as evenly distributed in a circle whereas $\epsilon = 1$ the momenta is distributed along an infinitely elongated object.

- Planar Flow: P_f [66]. Measures the degree to which the jets energy is evenly spaced over the plane defining the face of the jet. Planar flow is extracted by considering the following matrix,

$$I = \frac{1}{m_{\mathrm{jet}}} \sum_{i \in \mathrm{jet}} \frac{1}{E_i} \begin{pmatrix} (p_{x,i})^2 & (p_{x,i})(p_{y,i}) \\ (p_{y,i})(p_{x,i}) & (p_{y,i})^2 \end{pmatrix} \quad \ni \quad \hat{\boldsymbol{p}}_{\mathrm{jet}} \cdot \hat{\boldsymbol{z}} = 1 \ , \tag{4.5.8}$$

  where the components $p_{x,i}$, $p_{y,i}$ define the perpendicular momentum to the jet in each orthogonal direction. The eigenvalues of I, λ_1 and λ_2 determine the planar flow,

$$P_f = \frac{4\lambda_1\lambda_2}{(\lambda_1 + \lambda_2)^2} \ . \tag{4.5.9}$$

- Subjets: We can re-cluster the jet but now demanding that the algorithm truncate the procedure at N jets. These will define the N_{subjets} subjets of the original jet. These can be used to probe subjet structures explicitly:

  - N-subjettiness: τ_N [67]. A measure of to what degree a jet can be described

as N subjets.

$$\tau_N = \frac{1}{p_{\mathrm{T,\ jet}} R} \sum_{i \in \mathrm{jet}} p_{\mathrm{T},i} \min(\Delta R_{1,i}, \Delta R_{2,i}, ..., \Delta R_{N,i}) \ , \tag{4.5.10}$$

where each $\Delta R_{k,i}$ are calculated between the jet constituents i and candidate subjet, k up to N subjets. Values tending towards zero indicate a greater degree of subjettiness. These values on their own are not particularly insightful so we look at the ratios, $\tau_{xy} = \tau_x/\tau_y$ in particular τ_{21}.

- Rotation: θ_2. The angle required to align 2 subjets along a unit vector axis in $(\Delta\eta, \Delta\phi)$ with respect to the jet.
- Fractional energy of subjet: z_N. The fractional energy of the subjet with respect to the jet. We can also look at the energy balance between any 2 subjets, $z_{xy} = z_x/z_y$.
- Particle multiplicities: $N_{\mathrm{subjet,\ P}}$. Number of reconstructed particle tracks, can be all particles or subsets of charged ($N_{\mathrm{subjet,\ P^{\pm}}}$) or flavoured ($N_{\mathrm{subjet,\ P}^f}$). Which is not a collinear and infrared safe observable. This observable is not used in our cut-flow.
- Boosted plane angular separation: $\hat{\theta}_{\hat{\boldsymbol{n}},\hat{\boldsymbol{m}}}$. Boosting into the rest frame of the jet we can measure the angular separation between the subjets and an arbitrary plane defined by the unit vectors $\hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}$.
- Impact parameter: d_{k_i}. The impact parameter of the subjet i to the primary vertex.

4.5.3 Vertex Observables

Observables which characterise a vertex and the particles associated with it.

- Particle multiplicities: $N_{\mathrm{P}\in\mathrm{vertex}}$. Number of reconstructed particle tracks, can be all particles or subsets of charged ($N_{\mathrm{P}^{\pm}\in\mathrm{vertex}}$) or flavoured ($N_{\mathrm{P}^f\in\mathrm{vertex}}$).

Which is not a collinear and infrared safe observable. This observable is used indirectly by including particle level features of the hardest particles.

- Momentum: p^{μ}_{vertex}. Total four momentum flowing through the vertex. Similarly to the jets we can define further derived features:
  - Energy: $E_{\text{vertex}} = p^{0}_{\text{vertex}}$,
  - Transverse vertex momentum: $p_{\text{T, vertex}}$,
  - Invariant Mass: $m_{\text{vertex}} = \sqrt{p^{2}_{\text{vertex}}}$.
- Position: x^{μ}_{vertex}. The position of the reconstructed vertices can define the following:
  - Distance from primary vertex: $|\boldsymbol{x}_{\text{vertex}}|$. The physical distance from the primary vertex, can also express this in longitudinal and transverse distances, $|\boldsymbol{x}_{z,\text{vertex}}|$ and $|\boldsymbol{x}_{T,\text{vertex}}|$ respectively.
- Root mean square impact parameter: $d^{\text{RMSD}}_{\text{vertex}}$. Each track associated to each vertex has an impact parameter d_i, thus we can consider the spread of the track impact parameters.
- Order: $\mathcal{O}_{\text{vertex}}$. Tracking the momentum flow between vertices and their relative positions we can assign an ordering to their formation. Primary vertices are order 1, secondary order 2 and so on.

Particle multiplicities are discussed here as they feed into one of the ML architectures discussed in Chapter 5 indirectly. We take a selection of the most energetic particles associated to a vertex (as many as 5) and input their proprieties separately such that multiplicities inform the model parametrically. This large but by no means complete selection of observables provides a portal into the structure of different event topologies which we will use to build a cut-flow analysis and machine learning classifiers.

4.6 Analysis and Searches

This section will outline how a statistical search can be conducted using LHC data and standard statistical methods. Discovering or placing tighter constraints on measurements or parameters can be formulated as an hypothesis test. The methodology used in Chapter 5 uses the statistical tools and the CL_s method outlined in [7, 8].

After data on an event is collected we reconstruct the event into a list of final state particles where we know their momentum, charge and primary or secondary vertex. This information can be used to construct useful observables of an event. After many events we can construct distributions of many observables to conduct searches for new physics. A simple example of this are resonance searches. Any event that includes an unstable particle will follow a Breit-Wigner distribution and if the particle mass is within the range of centre of mass energies available to the collider the resonance will be accessible.

$$f(s) = \frac{1}{(s - m_R^2)^2 + \Gamma_R^2 m_R^2} \ , \tag{4.6.1}$$

where m_R is the mass of the resonance and Γ_R is the total decay width. The discovery of the Higgs boson followed this methodology, in Fig. 4.4 we see the invariant mass of the diphoton candidates histogrammed by number of events. We see a clear bump at $\sim$125 GeV, which is indicative of the Higgs.

To perform a statistical test we define a signal strength $\mu = \sigma/\sigma_{\text{SM}}$, which is the ratio of the measured cross section from a certain process from experimental or Monte Carlo data to the predicted cross section in the SM. The observable of the choice x is histogrammed into bins i, for example the invariant mass of diphotons as used in Fig. 4.4 for the determination of the Higgs mass. The data in each bin constitute the sum of all signal and background processes,

$$N_i = \mu s_i + b_i \ , \tag{4.6.2}$$

where if we expect a SM result $\mu \to 1$ and s_i and b_i are the number of signal and

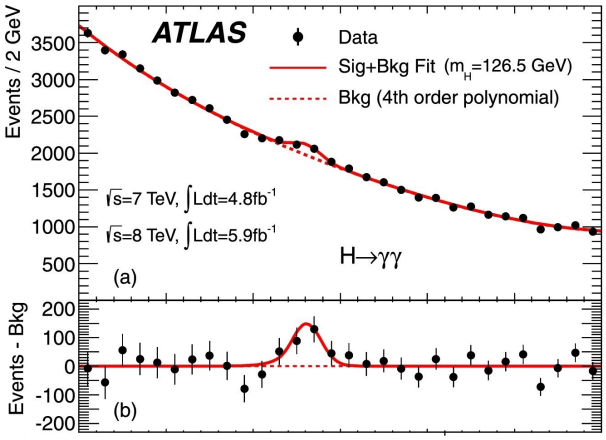


Figure 4.4: Distribution of invariant mass of diphoton candidates [52].

background events. There are two schemes determining the null hypothesis and alternative hypothesis:

- New signal discovery: $\mu \to 0$ where the background describing a known processes is our null hypothesis. The alternative hypothesis is when we have any quantity of signal processes $\mu \neq 0$.

- Limit setting: The null hypothesis is when $\mu \to 1$ describing the known background plus the BSM signal. This is tested against the background only alternative hypothesis, $\mu \to 0$.

We construct a likelihood function from the product of Poisson probabilities for each bin given n_i occurrences,

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \ . \tag{4.6.3}$$

$\boldsymbol{\theta}$ describes the set of nuisance parameters which characterise underlying probability distribution functions from which the signal and background distributions originate or alternatively, we add nuisance parameters which define uncertainty in measurements. For example, we can modify the background and signal bin values by Gaussian priors

to wrap in the associated errors,

$$\begin{aligned} s_i'(\boldsymbol{\theta}) &= \alpha_s s_i \ni \alpha_s \sim \mathcal{N}(\mu_{\alpha_s}, \sigma_{\alpha_s}) \ , \\ b_i'(\boldsymbol{\theta}) &= \alpha_b b_i \ni \alpha_b \sim \mathcal{N}(\mu_{\alpha_b}, \sigma_{\alpha_b}) \ . \end{aligned} \tag{4.6.4}$$

These normal priors $\mathcal{N}$ have to be included in the likelihood the profile likelihood ratio so that it is sensitive to changes in $\boldsymbol{\theta} = \{\mu_{\alpha_s}, \sigma_{\alpha_s}, \mu_{\alpha_b}, \sigma_{\alpha_b}\}$,

$$L(\mu, \boldsymbol{\theta}) \to \prod_{i=1}^{N} \frac{(\mu s_i' + b_i')^{n_i}}{n_i!} e^{-(\mu s_i' + b_i')} \mathcal{N}(\mu_{\alpha_s}, \sigma_{\alpha_s}) \mathcal{N}(\mu_{\alpha_b}, \sigma_{\alpha_b}) \ . \tag{4.6.5}$$

Now, to test a set of hypothesises with value μ we consider the profile likelihood ratio,

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \ . \tag{4.6.6}$$

The $\hat{\hat{\boldsymbol{\theta}}}$ is the value of $\boldsymbol{\theta}$ which maximizes L for a certain value of μ, it is the conditional maximum-likelihood estimator of $\boldsymbol{\theta}$. Whereas $\hat{\boldsymbol{\theta}}$ and $\hat{\mu}$ are the unconditional maximized estimators of L for which μ is allowed to vary freely. The introduction of nuisance parameters has the effect of broadening the likelihood function in μ compared to when the nuisance parameters are fixed, thus there is greater uncertainty and therefore loss of information due to the nuisance parameter uncertainties. Using $\lambda(\mu)$ we define a test statistic which measures the goodness of fit between the data and the hypothesized value of μ. The likelihood ratio varies between 0 and 1, where $\lambda = 1$ implies a good fit between data and a hypothesized value of μ. A standard choice for the test statistic is,

$$t_\mu = -2 \ln(\lambda(\mu)) \ . \tag{4.6.7}$$

As t_μ increases we have an decreasing compatibility of the data and μ. We quantify this discrepancy by computation of the p-value

$$p_\mu = \int_{t_{\mu,\text{obs}}}^{\infty} f(t_\mu|\mu) dt_\mu \ , \tag{4.6.8}$$

where $t_{\mu,\text{obs}}$ is the value of the test statistic from the observed experimental or Monte Carlo data. $f(t_\mu|\mu)$ defines the probability distribution function of t_μ for a

fixed signal strength μ. In the CL_s methodology we calculate p_μ in two cases, the background only assumption, CL_b and under the signal assumption as a function of μ, CL_{s+b}. Finally we have,

$$CL_s = \frac{CL_{s+b}}{1 - CL_b} = 1 - \alpha_{CL} \ , \tag{4.6.9}$$

where α_{CL} defines our desired confidence limit, if we wish to determine the 95% confidence limit we solve for μ in $CL_s = 0.05$. The value of $t_{\mu,obs}$ can either be determined experimentally, or by a simulated dataset. The probability distribution function, $f(t_\mu|\mu)$ can be determined from repeated sampling from the likelihood distribution and its associated priors. It is possible to create a direct relationship between μ and various SM coupling constants under the κ framework [68] or alternatively to a coupling constant corresponding to BSM physics.

Part II

Constraining Charming Higgs Decays

Chapter 5

Constraining Charming Higgs decays

5.1 Introduction

In this chapter we present work relating to the determination of improved limits on the charm Yukawa coupling. As the LHC moves into its high luminosity phase the precision of existing measurements will greatly improve. Couplings hitherto inaccessible will become subject to scrutiny providing exciting prospects for Higgs couplings to second generation quarks, rare decays or potential BSM physics [54]. In particular, I will present theoretical results for new confidence limits of the charm Yukawa coupling expressed under the κ framework for a LHC luminosity of $150\,\text{fb}^{-1}$ and $3000\,\text{fb}^{-1}$. This will be done by considering the following three channels: Vector Boson Fusion (VBF), W Higgs-Strahlung and Z Higgs-Strahlung production of a Higgs boson and its subsequent decay into charm quarks. These signal processes come with significant backgrounds so to eliminate these as much as possible and to reduce false positives, we apply a set of simple kinematic and jet feature cuts and build neural network architectures of three types: jet features, jet images and particle-level features. We then additionally consider two further neural network architectures with inputs involving vertex features and particle-level features to

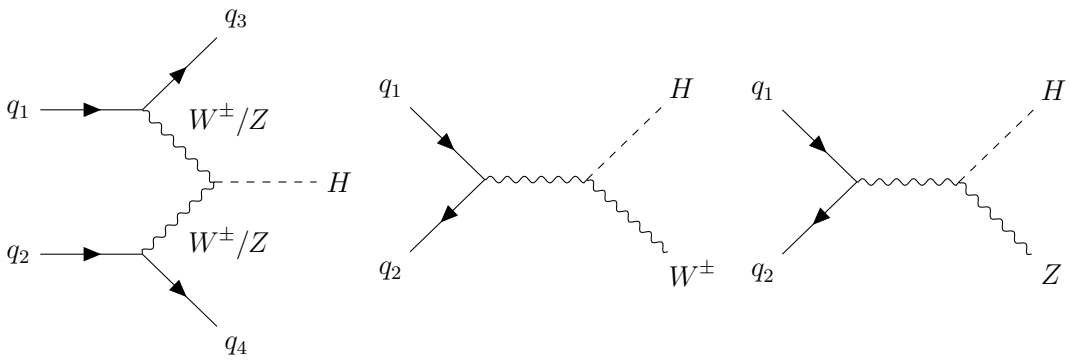
enhance light, charm and bottom jet discrimination.

The 2012 discovery of the Higgs boson H with a mass of 125 GeV by the ATLAS and CMS collaborations at the LHC [52, 53, 69] completed the experimental verification of the SM of particle physics and initiated a step-change in our quest to understand nature at its fundamental scale. Since this discovery, the focus has shifted to measuring the new bosons' properties and interactions and to constrain the effects of possible new physics manifesting through subtle deviations from SM predictions. A central part of these efforts has been the determination of the Higgs boson couplings to the other SM particles. By the end of Run 2, couplings to the SM gauge bosons [70–72] and to the massive third generation fermions [70, 73–78] have been measured, and the first coupling to the second-generation fermions, i.e. to the muon, has been constrained [70, 79, 80].

The next natural fermion coupling to consider is that of the Higgs Yukawa coupling to charm quarks which will provide a further, stringent test of the universal role of the Higgs boson in the generation of fundamental masses (see Section 4.1 and Appendix A). Various channels have been suggested, some of them introducing an associated photon [81], or the decay into $c\bar{c}$ quarkonia, for example in decays such as $H \to J/\psi + \gamma$ [82–84], or by recasting measurement of the $H \to b\bar{b}$ branching ratio, for example in [85–87]. We will analyse the prospects for such a measurement in three channels, namely the production of Higgs bosons and their subsequent decay into charm quarks in association with Z or W bosons and in VBF,

$$
\begin{aligned}
pp &\to H_{\to c\bar{c}}jj + X && \text{0-lepton channel, 0L. cf. Fig. 5.1, (left),} \\
pp &\to W_{\to \ell\nu_\ell} H_{\to c\bar{c}} + X && \text{1-lepton channel, 1L. cf. Fig. 5.1, (centre),} \\
pp &\to Z_{\to \ell\bar{\ell}} H_{\to c\bar{c}} + X && \text{2-lepton channel, 2L. cf. Fig. 5.1, (right).}
\end{aligned}
$$

In all processes it is important to consider the branching ratio of the Higgs boson to the charm quark. In the SM its value is $Br_{H\to c\bar{c}} = 2.89\%$ [70] is significantly smaller than its counterpart, the branching ratio to the b quarks, around 58.2% [70]. In previous studies [88, 89], the production of b quarks has been identified as a



VBF W Higgs-Strahlung Z Higgs-Strahlung

Figure 5.1: The feynman diagrams for the three Higgs production channels considered.

significant non-trivial background, along with signatures from QCD, vector boson and $t\bar{t}$ production.

To provide a simple interpretation of the results and facilitate straightforward comparison with experimental studies [88, 89], we will express the signal strength μ in the κ-framework [68] and use the $\mathrm{CL_s}$ method [8] to estimate the upper-bound for κ_c at the 95% level (cf. Section 4.6).

Modifying our observed signal counts, s_i by the signal strength μ_i in each bin, i of a given distribution,

$$N_i(\mu) = b_i + \mu s_i \,, \tag{5.1.1}$$

and assuming an unmodified background, $\mu = 1$, defines the SM value and variations in its value are equivalent to varying κ_c in some parametric way derived below. For different signal processes l, i.e. for the 0-lepton, 1-lepton and 2-lepton signatures, the μ_l represent ratios of cross sections σ_l times branching ratios in the narrow width approximation:

$$\mu_l = \frac{\sigma_l \cdot \mathrm{Br}_{H\to c\bar{c}}}{\sigma_l^{\mathrm{SM}} \cdot \mathrm{Br}_{H\to c\bar{c}}^{\mathrm{SM}}} \,, \tag{5.1.2}$$

where the σ_l denote the cross sections of the Higgs boson production processes l and $Br_{H\to c\bar{c}}$ is the branching ratio of the Higgs boson to charm quarks, $\Gamma^{\mathrm{Tot}}_{H\to c\bar{c}}/\Gamma_H$. The superscript "SM" indicates the SM values, its absence refers to values obtained

under variations of the signal strength. In the κ framework all couplings of particles X to the Higgs boson are independently modified by multiplying them with some κ_X thus transforming the total Higgs width,

$$\begin{aligned}\Gamma_H^{\mathrm{Tot}} = \Gamma_H^{\mathrm{Tot,SM}}\Big(&\kappa_s^2 Br^{\mathrm{SM}}_{H\to s\bar{s}} + \kappa_c^2 Br^{\mathrm{SM}}_{H\to c\bar{c}} + \kappa_b^2 Br^{\mathrm{SM}}_{H\to b\bar{b}} + \kappa_t^2 Br^{\mathrm{SM}}_{H\to t\bar{t}} + \kappa_\tau^2 Br^{\mathrm{SM}}_{H\to \tau\bar{\tau}} + \\ &\kappa_\mu^2 Br^{\mathrm{SM}}_{H\to \mu\bar{\mu}} + \kappa_Z^2 Br^{\mathrm{SM}}_{H\to \mathrm{ZZ}} + \kappa_W^2 Br^{\mathrm{SM}}_{H\to \mathrm{WW}} + \kappa_g^2 Br^{\mathrm{SM}}_{H\to gg} + \kappa_\gamma^2 Br^{\mathrm{SM}}_{H\to\gamma\gamma} + \\ &\kappa_{\gamma Z}^2 Br^{\mathrm{SM}}_{H\to\gamma Z}\Big)/\Big(1 - Br^{\mathrm{BSM}}\Big)\,.\end{aligned} \tag{5.1.3}$$

Where any decays beyond SM physics are represented by the BSM branching fraction Br^{BSM} which we set to 0. We also take the coupling to the first generation fermions to be negligible due to their very small masses compared with second- and third-generation particles. In the following, we will assume that only the charm Yukawa coupling is modified by a factor κ_c with all other κ_X set to 1, which in turn will result in modified partial and total decay widths of the Higgs boson and thus,

$$\mu_i = \left[\frac{\kappa_c^2 \Gamma^{\mathrm{SM}}_{H\to c\bar{c}}}{\Gamma_H^{\mathrm{Tot}}}\right] / \left[\frac{\Gamma^{\mathrm{SM}}_{H\to c\bar{c}}}{\Gamma_H^{\mathrm{Tot,\,SM}}}\right]\,. \tag{5.1.4}$$

In which the total width is simplified,

$$\Gamma_H^{\mathrm{Tot}} = \Gamma_H^{\mathrm{Tot,SM}} \kappa_c^2 Br^{\mathrm{SM}}_{H\to c\bar{c}}\,. \tag{5.1.5}$$

Previous analysis with such a framework sets the expected limit at 95% confidence for $\mu_{\mathrm{VH}(c\bar{c})} \leq 31^{+12}_{-8}$ at ATLAS [90] measured at an integrated luminosity of $139\,\mathrm{fb}^{-1}$ and $\mu_{\mathrm{VH}(c\bar{c})} \leq 7.6^{+3.4}_{-2.3}$) at CMS [91] measured at $138\,\mathrm{fb}^{-1}$.

The quadratic dependence of κ_c on μ leads to some limitations in the maximal resolvable value of μ that leads to a meaningful value of κ_c. This value sits at $\mu = 1/Br^{\mathrm{SM}}_{H\to c\bar{c}} = 34.6$.

5.2 Simulation

For the analysis, signal and background samples are generated with SHERPA 2.2.9 [92], using LO-merged samples throughout [93–95]. The signal and background processes considered are:

- VBF, with $H \to c\bar{c}$ (signal class 0L) and $H \to b\bar{b}$,
- WH-associated production, with $H \to c\bar{c}$ (signal class 1L) and $H \to b\bar{b}$,
- ZH-associated production, with $H \to c\bar{c}$ (signal class 2L) and $H \to b\bar{b}$,
- vector boson ($W \to \ell\nu$, $Z \to \ell\bar{\ell}$) production + jets,
- vector boson pair production + jets, where one of the two bosons decays hadronically,
- pure QCD multijets,
- $t\bar{t}$ production + jets,

where we merge up to two additional jets to the core process. We use the NNPDF 3.0 PDFs [30] from LHAPDF [31], the COMIX matrix element generator [28] for the LO matrix element, the CSSHOWER [33] for the simulation of QCD radiation, AHADIC++ as the hadronisation model [96], PHOTONS++ [97] for the emission of photons in the decays of the W and Z bosons and SHERPA's built-in models for the underlying event and hadron decays. Using the default prescription for setting the renormalisation and factorisation scales in multi-jet merging, we obtain theoretical uncertainties from their variation by a factor of $f_{R,F} = 2$ in both directions and forming the envelope of the 7-points, schematically

$$\{f_R,\, f_F\} = \left\{(\frac{1}{2},\, \frac{1}{2})\ ;\ (\frac{1}{2},\, 1)\ ;\ (1,\, \frac{1}{2})\ ;\ (1,\, 1)\ ;\ (1,\, 2)\ ;\ (2,\, 1)\ ;\ (2,\, 2)\right\}. \quad (5.2.1)$$

The rationale for using this approach is two-fold: First of all, including higher-order QCD corrections does not induce any sizable change in the shape of distributions, but

only alters the overall cross section [98] which can usually be captured by applying a flat K-factor to the overall sample. For the processes we consider, this K-factor is of the order of 1.3 or below, thereby increasing the total number of events by up to 30%, which in turn translates to a decrease of the statistical uncertainty by about 10%. Secondly, apart from increasing total event numbers, the higher-order corrections reduce scale uncertainties, typically by a factor of two or more. As we are only able to roughly estimate experimental uncertainties, our approach of potentially overestimating the theory uncertainties therefore merely translates into the results being more conservative.

5.3 Analysis Strategy

5.3.1 Initial Cuts

For each of the three signal topologies, there is a unique set of cuts summarised in Tab. 5.1. These cuts are chosen to take advantage of the event shape and jet structures provided by the Higgs fatjet structure and the isolated lepton properties. The observables used in these cuts are presented in Section 4.5 and details of the jet clustering algorithm and its parameters in Section 4.4. They are encoded in a RIVET [99] analysis and detailed by:

1. $E_{\mathrm{T,\,miss}}$ is reconstructed from the total sum of visible particles i, with

$$|\eta_i| < 4 \text{ and } p_{T,i} > 100\,\mathrm{MeV}\,, \tag{5.3.1}$$

   and it is particularly powerful in enhancing or suppressing events with (1L) and without (0L, 2L) decaying W bosons.

2. To isolate leptons, we demand that the total transverse energy of all particles in a cone of angular size $R_{\mathrm{iso}} = 0.2$ (cf. (4.2.6)) around the lepton direction is

Cut	#	0L	1L	2L
$E_{\mathrm{T,\,miss}} \leq 30\,\mathrm{GeV}$	1	✓	X	✓
$E_{\mathrm{T,\,miss}} \geq 30\,\mathrm{GeV}$	1	X	✓	X
0 Isolated Leptons	2	✓	X	X
1 Isolated Leptons	2	X	✓	X
2 Isolated Leptons	2	X	X	✓
1+ fatjet	3	✓	✓	✓
Candidate fatjet	4	✓	✓	✓
2 forward QCD jets	5	✓	X	X
$p_{\mathrm{T,\,miss}} + p_{\mathrm{T}\,\ell_{\mathrm{iso},1}}$ and fatjet B2B	5	X	✓	X
$p_{\mathrm{T}\,\ell_{\mathrm{iso},1}} + p_{\mathrm{T}\,\ell_{\mathrm{iso},2}}$ and fatjet B2B	5	X	X	✓
1+ Secondary vertices	6	✓	✓	✓
2 subjets	7	✓	✓	✓
Simple Vertex cuts	8	✓	✓	✓
Machine Learning cuts	9	✓	✓	✓

Table 5.1: Cut-flow for each channel.

constrained by 5% of the lepton transverse energy $E_{T,\ell}$,

$$\sum_{\Delta R_i < R_{\mathrm{iso}}} E_{T,i} \;\leq\; 0.05 \cdot E_{T,\ell}\;. \tag{5.3.2}$$

We require the exact number of 0, 1 ($\ell_{\mathrm{iso},1}$), or 2 ($\ell_{\mathrm{iso},1}$, $\ell_{\mathrm{iso},2}$) isolated leptons for the 0L, 1L and 2L topologies respectively.

3. We demand the Higgs decay products to form a fatjet, defined by the anti-k_{T} algorithm [62, 100] with $R = 1.0$ and $p_{\mathrm{T}} > 250\,\mathrm{GeV}$ and we require events to contain at least one such fatjet.

4. To identify the required single candidate fatjet, the highest-p_T fatjet must contain at least three particles, but no isolated lepton and its invariant mass must satisfy

$$75\,\mathrm{GeV} < m_{\mathrm{jet}} < 175\,\mathrm{GeV}\,, \tag{5.3.3}$$

cf. Fig. 5.2 for an illustration that motivates our choice.

5. In addition, we place some cuts that uniquely identify specific signal topologies:

   - for the VBF (0L) topology, we require two forward anti-k_{T} jets ($R = 0.4$,

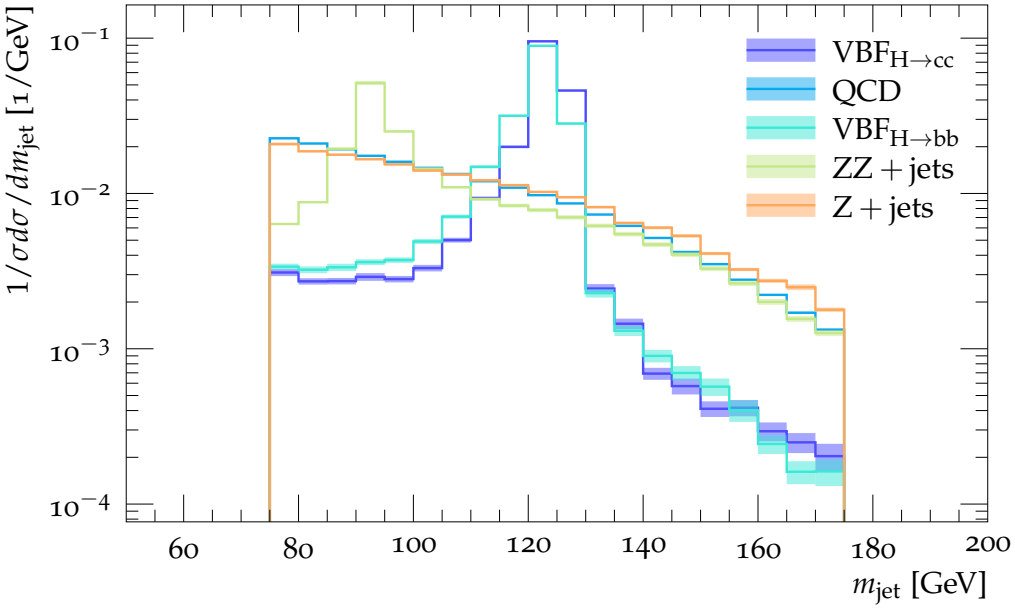


Figure 5.2: Normalised histogram of the reconstructed fatjet mass in the 0L channel. Shaded region indicates the error estimated from PDF4LHC scale variation convention and statistics.

$p_{\text{T}} > 20\,\text{GeV}$) j_1 and j_2 and with a minimal rapidity separation and combined invariant mass,

$$\Delta y_{j_1 j_2} > 2.5 \quad \text{and} \quad m_{j_1 j_2} > 400\,\text{GeV}\,. \tag{5.3.4}$$

- For the two VH topologies (1L and 2L), we demand that the combined momentum of the two isolated leptons (2L) or of the isolated lepton and missing transverse momentum (1L) is anti-parallel to the fatjet momentum within a tolerance of $R = 0.4$.

6. The fatjet must contain at least one reconstructed secondary vertex with at least two charged tracks. An adapted vertex fitter [57, 101] (see Section 4.3) performs a minimisation of weighted impact parameters, d_i over points of closest approach of charged particles contained within the fatjet. Vertices within 1 mm of each other are considered unresolved and are merged.

7. The fatjet must contain at least two subjets (anti-k_{T}: $\Delta R = 0.4$, $p_{\text{T}} > 20\,\text{GeV}$).

8. Further cuts are applied on the reconstructed primary vertex, namely the invariant mass of the total momentum of particles associated to it and the root mean square distance of particle tracks to the vertex,

$$m_{\text{vertex}} < 1\,\text{TeV} \quad \text{and} \quad d^{\text{RMSD}}_{\text{primary}} < 3\,\text{mm}\,, \tag{5.3.5}$$

which are determined empirically studying histograms of distributions over signal and background.

In Fig. 5.3 we exhibit, as an example, the resulting cut-flow for the 0L channel (VBF).

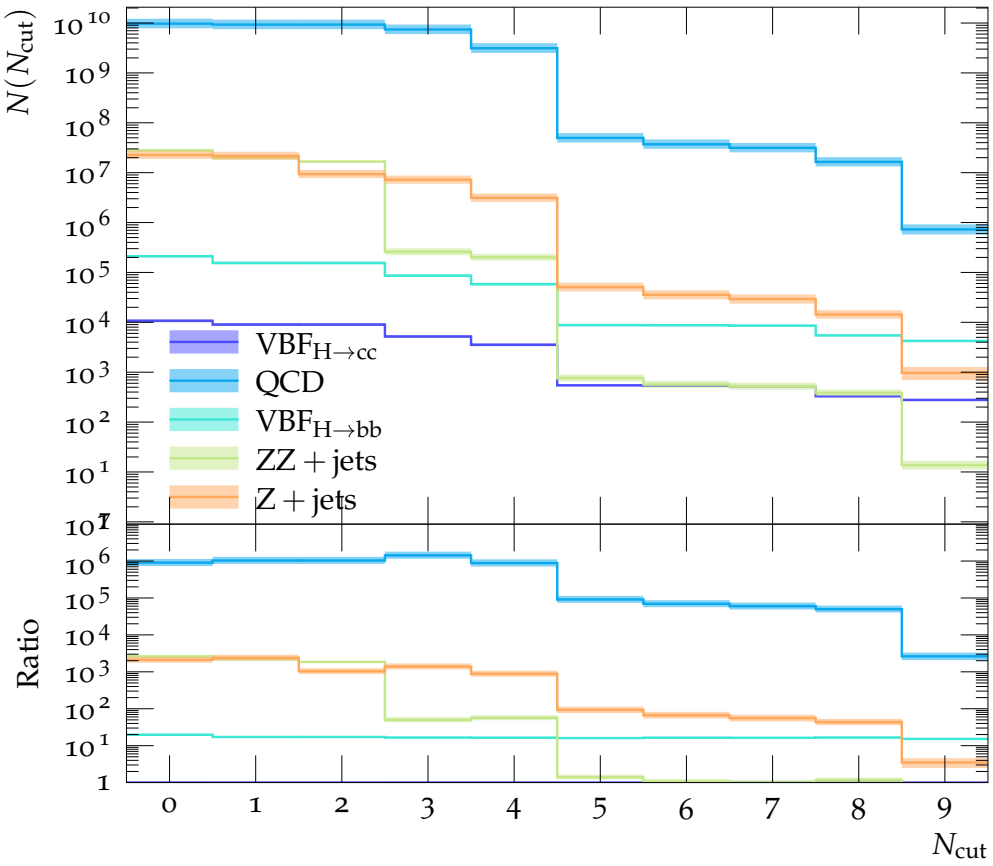


Figure 5.3: Example cut-flow for the 0L channel. The N^{th} cut, N_{cut} is performed in the N^{th} bin of the x-axis. Shaded region indicates the error estimated from PDF4LHC scale variation convention and statistics. Backgrounds where $N_{\text{process}} << N_{\text{VBF}_{H\to c\bar{c}}}$ are omitted for clarity.

5.4 Machine Learning improvements

5.4.1 ML "Booster"

In a ninth stage of the cut-flow, we boost the cut-based analysis through a set of multivariate neural networks (MVA) with combined inputs from three different data types. The input features used in these cuts are presented in Section 4.5. This MVA is trained with TensorFlow [102] on $\sim$10,000 events of each background and signal processes that have survived the initial cuts:

1. "Observable": a dense fully connected network trained with global event and fatjet features. A large selection of features are fed into the multivariate neural network. A number of global event and jet observables can be considered[1] in order to best distinguish between the classes and only the features with the strongest Shapley values [103] (see Fig. 5.5 – Fig. 5.7) or uniqueness with respect to the other MVA input datatypes. Shapley values provide a quantitative measure of cooperation between different features in a ML method. Specifically we compare the mean Shapley contribution across all classes, $\bar{S}$ which provides a metric of the classifying power of a feature across all classes. Features seen to be contributing the least to classification can be removed from the model, which decreases the model complexity with little loss in classifying power. The chosen subset of observables used are:

    - jet mass: m_{jet},
    - missing transverse energy: $E_{\text{T, miss}}$,
    - total perpendicular momentum in jet: $p_{\perp \in \text{jet}}$,
    - 2-subjettiness: τ_2, subjets are re-clustered by the k_T algorithm with $R = 1.0$.
    - subjet energy fraction: z_1 and z_2,

[1] A full list of the considered observables and their definitions can be found in Section 4.5.

- planar flow: P_f.

These observables are standardised and normalised,

$$O'_i = \frac{O_i - \mu(O)}{\sigma(O)} \,. \tag{5.4.1}$$

This normalisation is a standard ML technique to improve training and performance (cf. Section 2.2).

2. "Image": a dense fully connected convolutional neural network trained on rotated "jet images". We create "two-dimensional calorimeter" images for the fatjet centred on its axis and apply simple pre-processing steps to standardise the images using standard computer vision techniques (see Section 2.2.2):

   (a) Center: Redefine co-ordinates in (η, ϕ) such that the jet axis lies at (0,0).

   (b) Rotate: Rotate (η, ϕ) such that any subjets align on the ϕ-axis[1].

   (c) Image set-up: An 'image' which spans $\eta, \phi \in (-R, R)$ with 21x21 pixels.

   (d) Build: For each particle, i in the jet add some variable x in the bin (η_i, ϕ_i).

   (e) Scale: For each image scale such that $0 < I^{\eta,\phi} < 255$.

The algorithm above can be extended from grey-scale to a colour image, in this work we consider RGB images where the pixel variable x is,

$$R = -\sum_{\mathrm{P}\in\mathrm{jet}} \log\left(E_i/E_{\mathrm{jet}}\right) \,,$$
$$G = -\sum_{\mathrm{P}\in\mathrm{jet}} \log\left(p_{\mathrm{T,\,jet}}\right) \,,$$
$$B = N_{\mathrm{P}^{\pm}\in\mathrm{jet}} \,.$$

Jet images exemplified in Fig. 5.4 are fed into the CNN. A pre-processing function centres and normalises the images in each colour channel c, I_c such

[1] This defines the observable θ_2.

that they only contain integer values between 0-255,

$$I_c^{\eta_i,\phi_i} = \frac{I_c^{\eta_i,\phi_i} - \min(I_c^{\eta,\phi})}{\max(I_c^{\eta,\phi}) - \min(I_c^{\eta,\phi})} \,. \tag{5.4.2}$$

Similarly to the observables dataset logarithm and normalisation improves the gradient decent performance, allowing the network to be more sensitive to the full range of feature values and enables faster learning during backpropagation. The Image datasets also undergo some on-the-fly augmentation; once imported, there is a random chance of pixel shifting and randomised horizontal and vertical flipping,

- Flipping: $I_c^{\eta,\phi} \to I_c^{-\eta,\phi}$, $I_c^{\eta,-\phi}$, $I_c^{-\eta,-\phi}$.
- Shifting: $I_c^{\eta,\phi} \to I_c^{\eta\pm\Delta,\phi}$, $I_c^{\eta,\phi\pm\Delta}$, $I_c^{\eta\pm\Delta,\phi\pm\Delta}$.

These changes artificially generate more data and enforce discrete symmetries and any make the network robust to any numerical issues encountered.

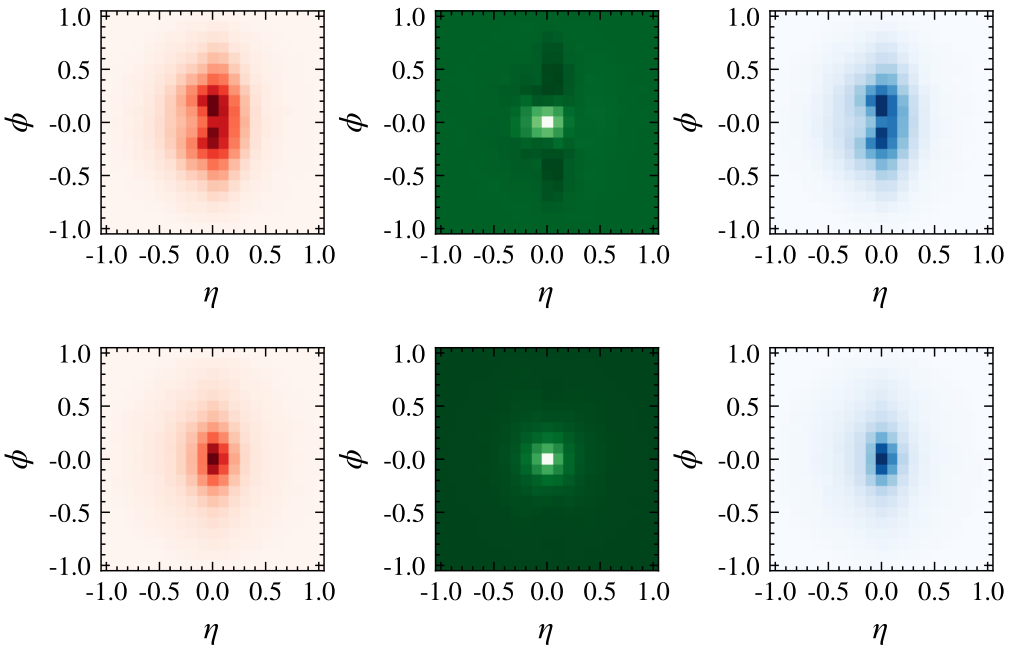


Figure 5.4: Example mean fatjet images signal and background in the 0L, VBF channel for each colour channel, Left: R. Center: G. Right: B. Top: VBF($H_{\to c\bar{c}}$) Bottom: QCD.

3. "Flow": a dense fully connected recursive neural network trained on ordered particle-level features within the fatjet. particle-level features are fed into a recursive neural network. To provide the neural network with a structured sequence of particles, up to ten fatjet constituents are ordered in energy. The particle-level features are:

   - η displacement w.r.t. fatjet axis: $\Delta\eta_p = \eta_p - \eta_{\text{jet}}$.
   - ϕ displacement w.r.t. fatjet axis: $\Delta\phi_p = \phi_p - \phi_{\text{jet}}$.
   - R displacement w.r.t. fatjet axis, $\Delta R_p = \sqrt{\Delta\eta_p^2 + \Delta\phi_p^2}$.
   - Energy: $\log\Big(E_p\Big)$.
   - Perpendicular momentum: $\log\Big(p_{\text{T},p}\Big)$.
   - Energy fraction: $\log\left(\dfrac{E_p}{E_{\text{jet}}}\right)$.
   - Perpendicular momentum fraction: $\log\left(\dfrac{p_{\text{T},p}}{p_{\text{T, jet}}}\right)$.

   Again we standardise these particle features across the training dataset and all particles,

$$F_i' = \frac{F_i - \mu(F)}{\sigma(F)} \ . \tag{5.4.3}$$

The structure of each of these parts are shown in Fig. 5.8.

The neural networks are trained over many epochs with a train-validation split of 90%–10% and the network with the highest validation accuracy from any epoch is kept. We note that at this stage there is confusion between the $H \to c\bar{c}$ and $H \to b\bar{b}$ class as this ML "booster" makes no attempt to build a b and c jet classifier. The distinction between $H \to b\bar{b}$ and $H \to c\bar{c}$ is addressed with another network architecture, the ML charm-bottom discriminator. In each channel we report overall retention of signal excluding $H \to b\bar{b}$. The results for signal retention and background rejection are summarised in Tab. 5.2 for ϵ_s and ϵ_b the signal acceptance and background rejection efficiencies (excluding $H \to b\bar{b}$), where the cut is performed by only retaining events in which the signal class has the highest predictive probability in the network.

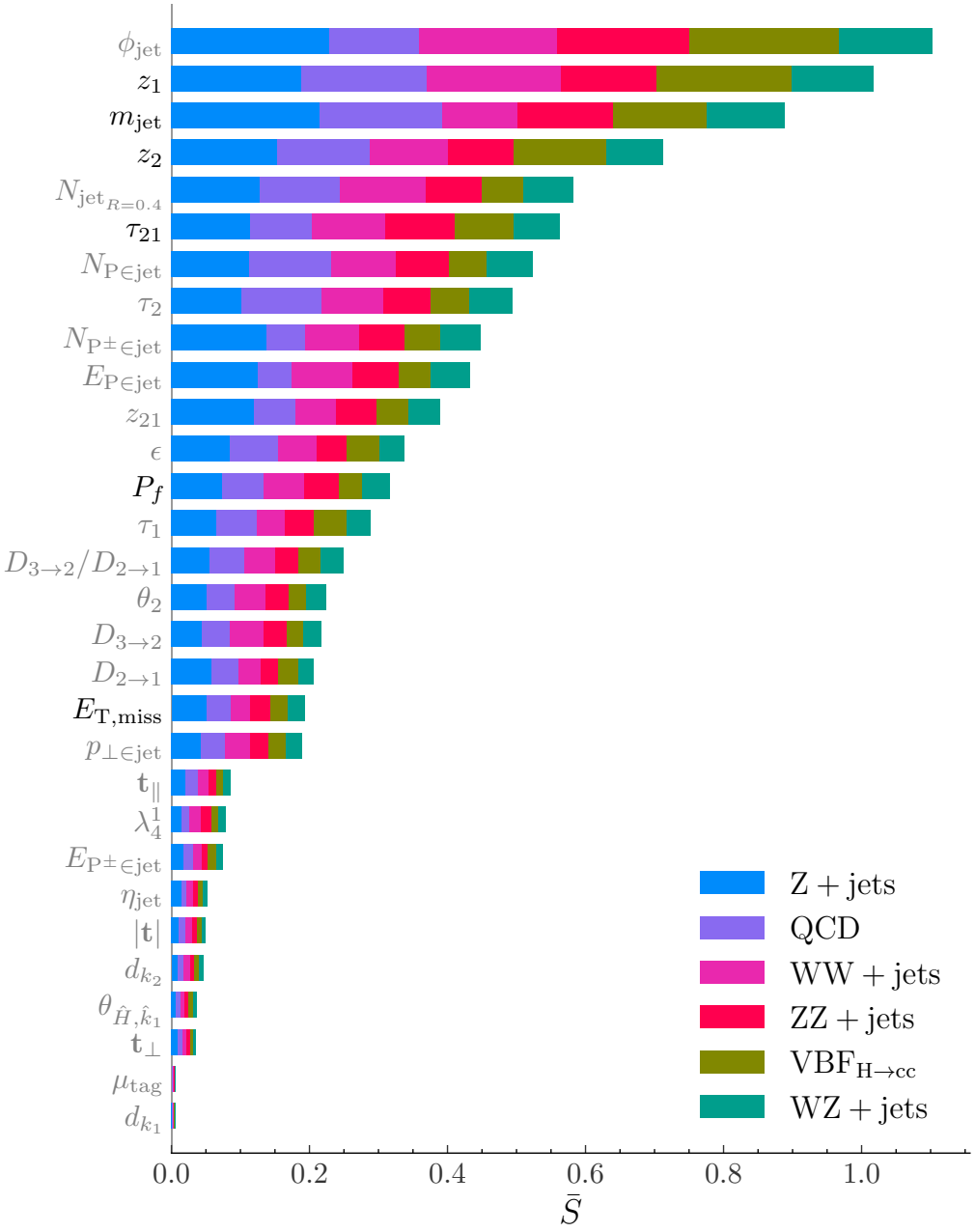


Figure 5.5: Absolute Shapley contributions to classification for a FCDN trained with all observables in the 0L, Vector Boson Fusion channel. The final choice of observables for the ML "booster" network are indicated in black coloured text , others in grey. The Shapley contribution to each channel classification of dominant processes are shown in different colours.

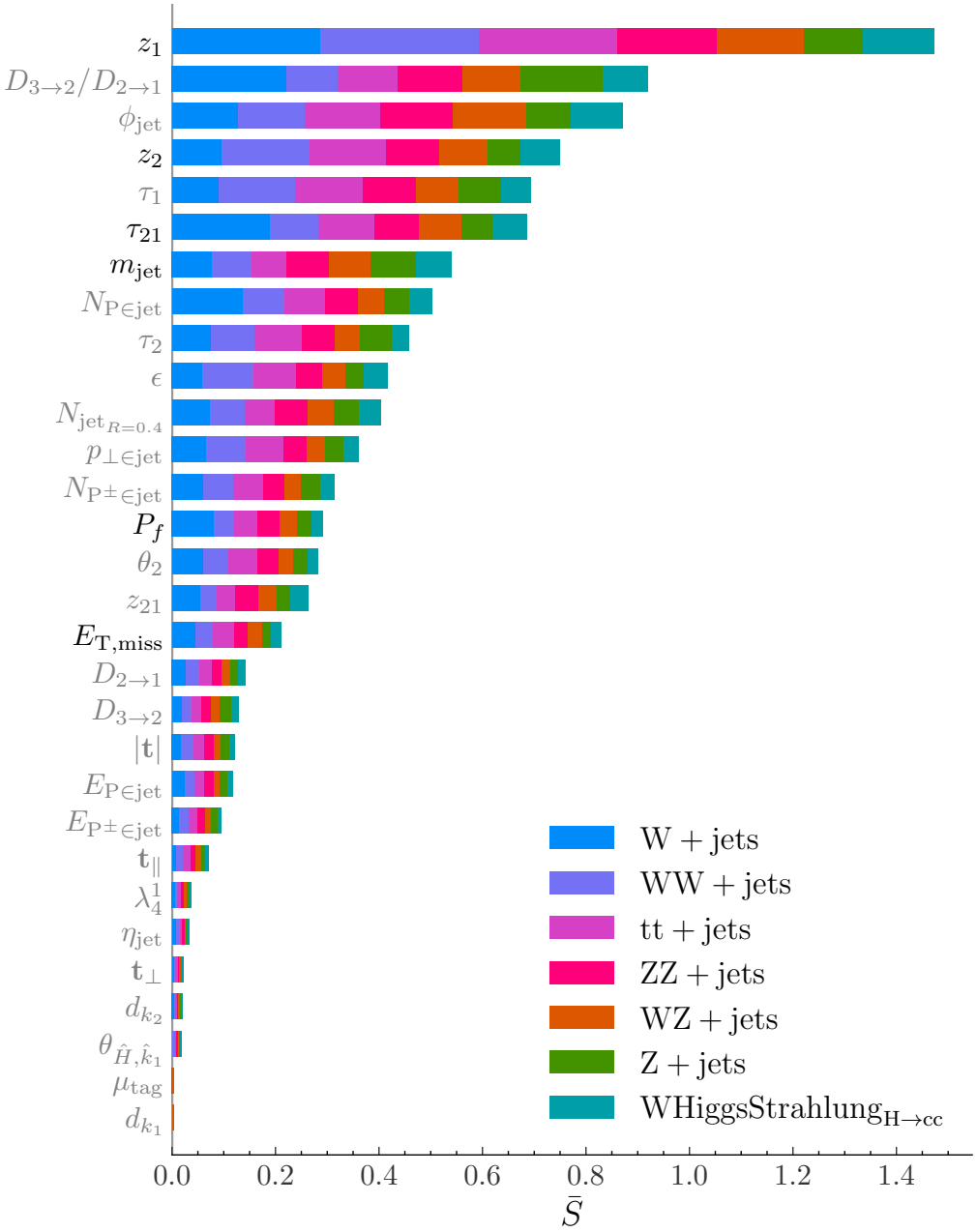


Figure 5.6: Absolute Shapley contributions to classification for a FCDN trained with all observables in the 1L, W Higgs-strahlung channel. The final choice of observables for the ML "booster" network are indicated in black coloured text , others in grey. The Shapley contribution to each channel classification of dominant processes are shown in different colours.

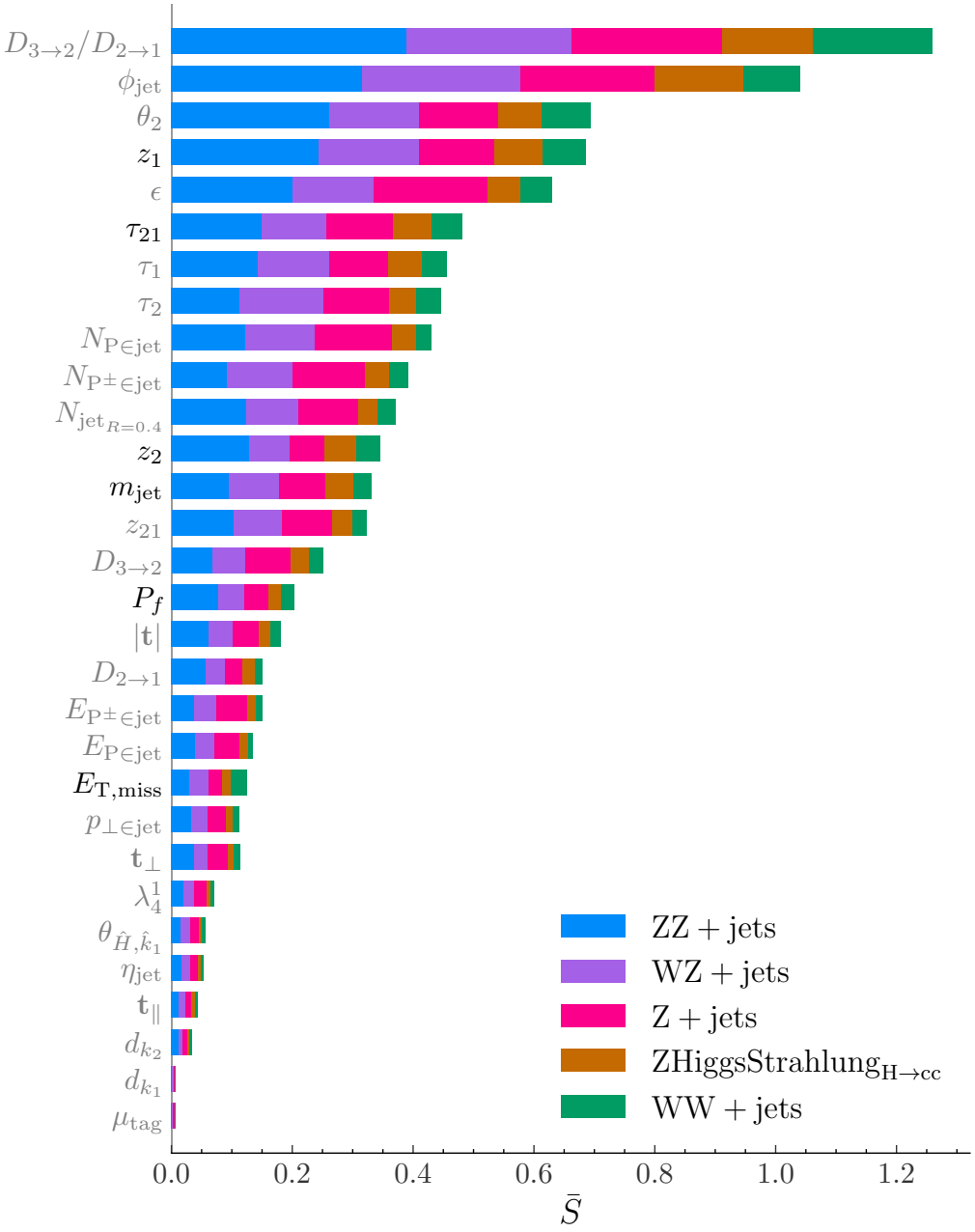


Figure 5.7: Absolute Shapley contributions to classification for a FCDN trained with all observables in the 2L, Z Higgs-strahlung channel. The final choice of observables for the ML "booster" network are indicated in black coloured text , others in grey. The Shapley contribution to each channel classification of dominant processes are shown in different colours.

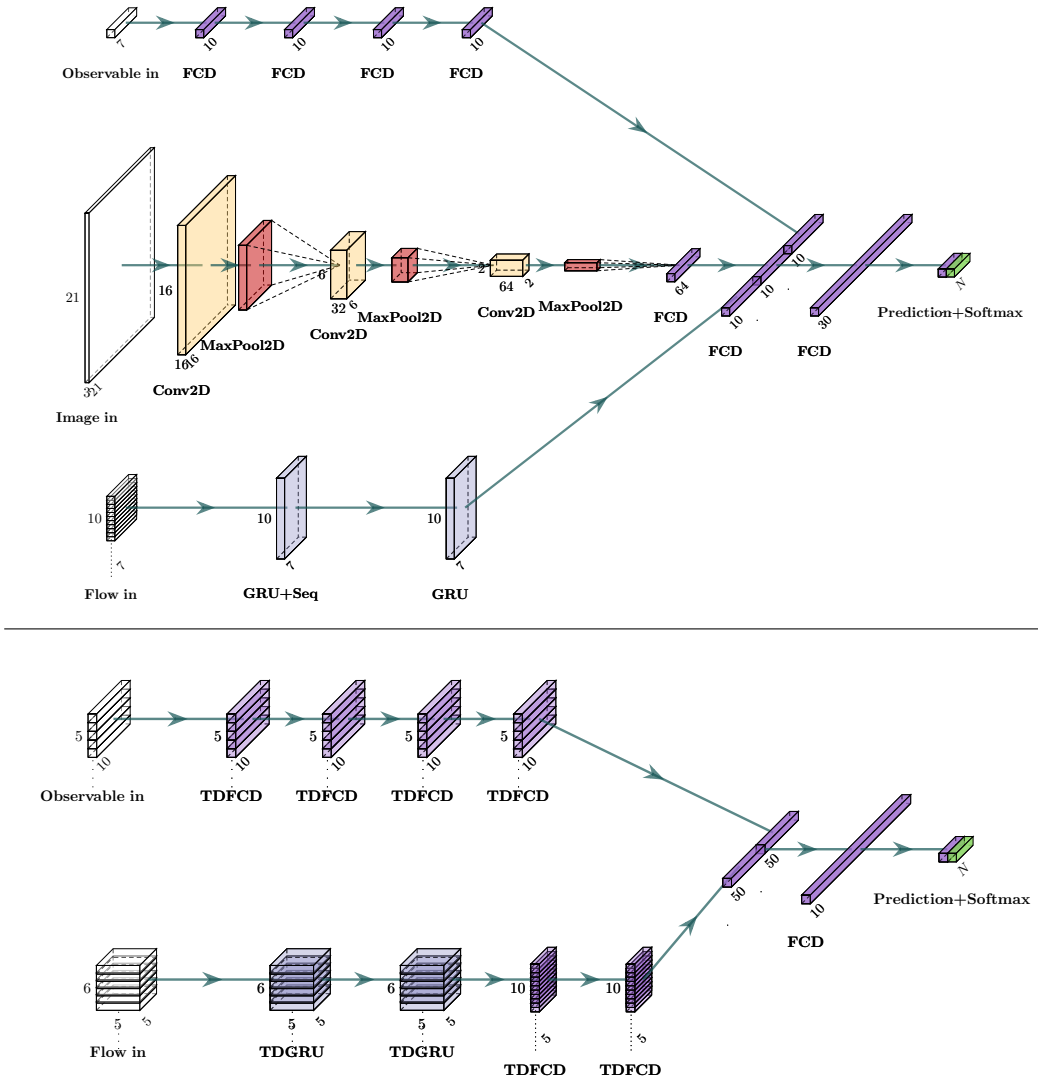


Figure 5.8: The neural network architectures; Top: $ML_{booster}$. Bottom: $ML_{b\leftrightarrow c}$.
The abbreviation layer names are; **FCD**: Fully connected dense layer, **Conv2D**: A two dimensional convolutional layer, **MaxPool2D**: A two dimensional maximum pooling layer, **GRU+Seq**: A Gated Recurrent Unit which returns state from each unit, **GRU**: A Gated Recurrent unit, **TDFCD**: Time distributed fully connected dense layer and lastly **TDGRU**: Time distributed gated recurrent unit. Drop out layers used in training are omitted from these diagrams.

NN	#L	ϵ_s	ϵ_b
MVA	0	83.9%	78.1%
MVA	1	72.0%	84.7%
MVA	2	81.5%	74.7%

Table 5.2: $\text{ML}_{\text{booster}}$ Neural network efficiencies for each channel.

The trained models are converted into a format suitable to run natively in RIVET [99] using the Frugally Deep header library [104].

5.4.2 ML Charm vs. Bottom Discriminator

Finally, we construct a neural network to discriminate the $H \to c\bar{c}$ signal from the $H \to b\bar{b}$ background, based on the structure of the displaced vertices. There are many examples of superior classifiers for c and b jet classifiers used by experiments, including MV2 and DL1 [105]. However, here we construct our own multivariate neural network that uses primary and secondary vertex features to discriminate between fatjets with only light constituents, with c hadrons and with b hadrons. This network has two inputs:

1. "vertex observable": a time distributed fully connected network with input features describing up-to five reconstructed vertices with ten features:
    - number of reconstructed vertices: N_{vertex}.
    - total number of tracks: $N_{\text{P}\in\text{vertex}}$.
    - vertex invariant mass: m_{vertex}.
    - vertex energy: E_{vertex}.
    - distance from primary vertex: $|x_{\text{vertex}}|$.
    - transverse distance from primary vertex: $|x_{\text{T, vertex}}|$.
    - RMSD of impact parameters: $d^{\text{RMSD}}_{\text{vertex}}$.

- standard deviation of squared impact parameters: $\sigma(d^2_{\text{vertex}})$.
- polar angle of vertex: θ_{vertex}.
- order of vertex: $\mathcal{O}_{\text{vertex}}$.

$\mathcal{O}_{V_i}$ is necessarily 0 for the primary vertex; then any vertices within a cone with opening angle $\theta = \pi/4$ are subsequently numbered in order of distance from the primary vertex. This provides the neural network with reinforcement of a natural ordering in displaced vertices in any event.

2. "vertex flow": uses particle-level features of the five hardest particles, p of each vertex with the following inputs:

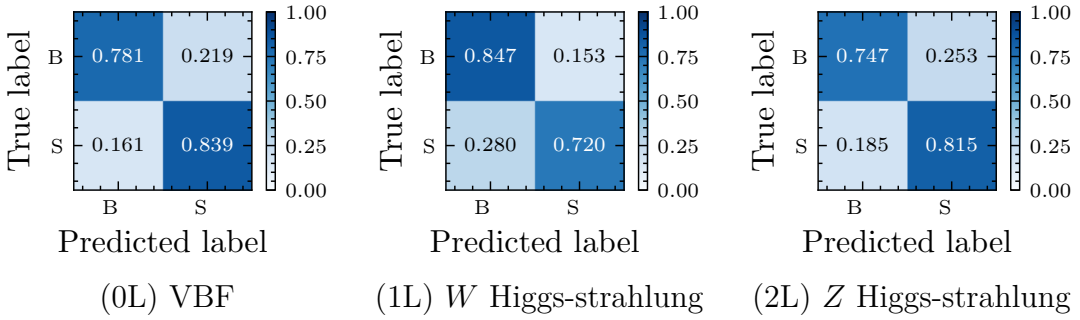
- η displacement w.r.t. fatjet axis: $\Delta\eta_p = \eta_p - \eta_{\text{jet}}$.
- ϕ displacement w.r.t. fatjet axis: $\Delta\phi_p = \phi_p - \phi_{\text{jet}}$.
- R displacement w.r.t. fatjet axis, $\Delta R_p = \sqrt{\Delta\eta_p^2 + \Delta\phi_p^2}$.
- energy fraction: $\log\left(\frac{E_p}{E_{\text{jet}}}\right)$.
- longitudinal impact parameter: $d_{z,p}$.
- transverse impact parameter: $d_{\text{T},p}$.

All of these features are normalised over a weighted mean over all features for all classes, leading to the results summarised in Tab. 5.3. The vertex booster network

NN	#L	ϵ_s	ϵ_b
Vertex MVA	0	58.8%	84.3%
Vertex MVA	1	66.2%	78.5%
Vertex MVA	2	51.3%	90.0%

Table 5.3: $\text{ML}_{b\leftrightarrow c}$ Neural Network efficiencies for each channel.

was also independently trained on a streamlined dataset consisting of $H \to b\bar{b}$, $H \to c\bar{c}$ and QCD fatjets with a minimal cut-flow for comparison to other analyses. We find $\epsilon_{H\to c\bar{c}} = 72\%$ and background rejection $\epsilon_b = 75\%$. Comparing directly with



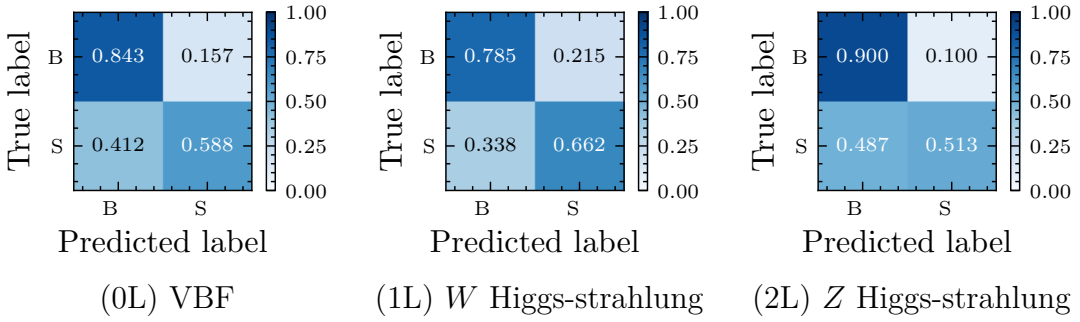


Figure 5.9: Normalised contact matrices for signal versus combined backgrounds for each channel; Top: $\mathrm{ML}_{\mathrm{booster}}$. Bottom: $\mathrm{ML}_{\mathrm{b}\leftrightarrow\mathrm{c}}$.

the JetFitterCharm Algorithm [106] and demanding similar signal efficiencies, we obtain background rejection rates summarised in Tab. 5.4.

	ϵ_c	$1/\epsilon_b$	$1/\epsilon_l$
"Loose"	0.95	1.65	1.03
"Medium"	0.21	13.2	149

Table 5.4: Summary of neural network efficiencies selected ϵ_c on the streamlined fatjet dataset, resulting in ϵ_l light jet, ϵ_b bottom jet and ϵ_c charm jet efficiencies.

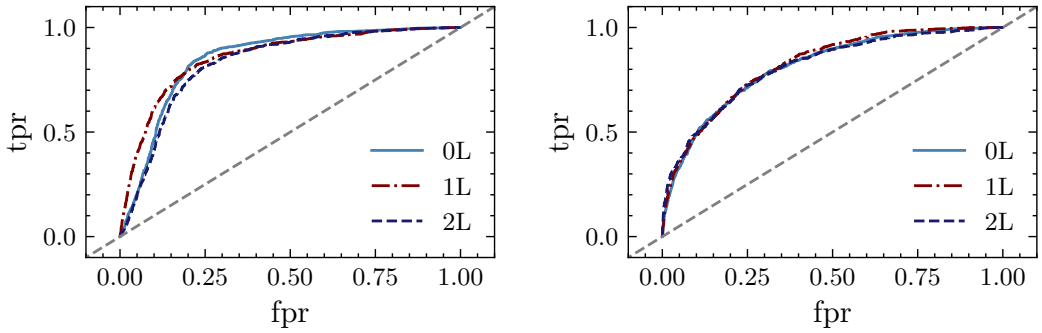


Figure 5.10: ROC curves for each channel and architecture; Left: $\mathrm{ML}_{\mathrm{booster}}$. Right: $\mathrm{ML}_{\mathrm{b\leftrightarrow c}}$. We report AUC of 0.85, 0.86, 0.83 for channel 0, 1 and 2 respectively for the $\mathrm{ML}_{\mathrm{booster}}$ and then 0.81, 0.82, 0.81 for channel 0, 1 and 2 respectively for $\mathrm{ML}_{b\leftrightarrow c}$ for the correct classification of only our signal process in each channel.

5.5 Results

5.5.1 Limitations of the κ Framework

To determine the 95% confidence limit on the signal strength, μ_{0L}, μ_{1L} and μ_{2L} we use a $\mathrm{CL_s}$ [8] frequentist approach implemented in RooFit/RooStats [107–109] and treat SHERPA as a SM simulator (see Section 4.6). $\mu(\kappa, Br)$ is derived from Eq. (5.1.4) as

$$\mu_c = \frac{\kappa_c^2}{1 + Br^{\mathrm{SM}}_{H\to c\bar{c}}(\kappa_c^2 - 1)} \,. \tag{5.5.1}$$

We use the $\mathrm{CL_s}$ method to determine the confidence limits on μ_i. The likelihood function used incorporates uncertainties due to statistics $\sigma_{N,x} = \sqrt{N}$, luminosity $\sigma_L = 2.5\%$ [110] and scale variations σ_α. Therefore,

$$\begin{aligned}\mathcal{L}(\mu, \mathbf{s}, \mathbf{b}) = \mathcal{N}(L', L, \sigma_L)\mathcal{N}(\alpha'_s, \alpha_s, \sigma_{\alpha_s})\mathcal{N}(\alpha'_b, \alpha_b, \sigma_{\alpha_b}) \\ \times \prod_x \mathcal{P}(b_x + s_x, L'(\alpha'_b b'_x + \mu\alpha'_s s'_x))\mathcal{N}(b'_x, b_x)\mathcal{N}(s'_x, s_x) \,.\end{aligned} \tag{5.5.2}$$

The uncertainties are parameterised with a Gaussian smearing over our expected

values and the priors $\mathcal{N}$ and $\mathcal{P}$ are Gaussian and Poisson distributions, respectively. Profiling the likelihood function for each channel determines the confidence limits as a function of μ_i, with $i = 0\text{L}, 1\text{L}$ or 2L. These independent channels are combined into one confidence limit which could be inverted to a confidence limit on κ_c. However, when inverting Eq. (5.5.1) we get,

$$\kappa_c^2 = \frac{\mu_i(1 - Br^{\text{SM}}_{H\to c\bar{c}})}{1 - \mu Br_{H\to c\bar{c}}} \,, \tag{5.5.3}$$

which does not have a well defined value of κ_c for $\mu_c > 1/Br_{H\to c\bar{c}}$.

As we explore this region of μ-values with our analysis and in order to avoid counter-intuitive results for κ_c, we only quote projections for limits on μ.

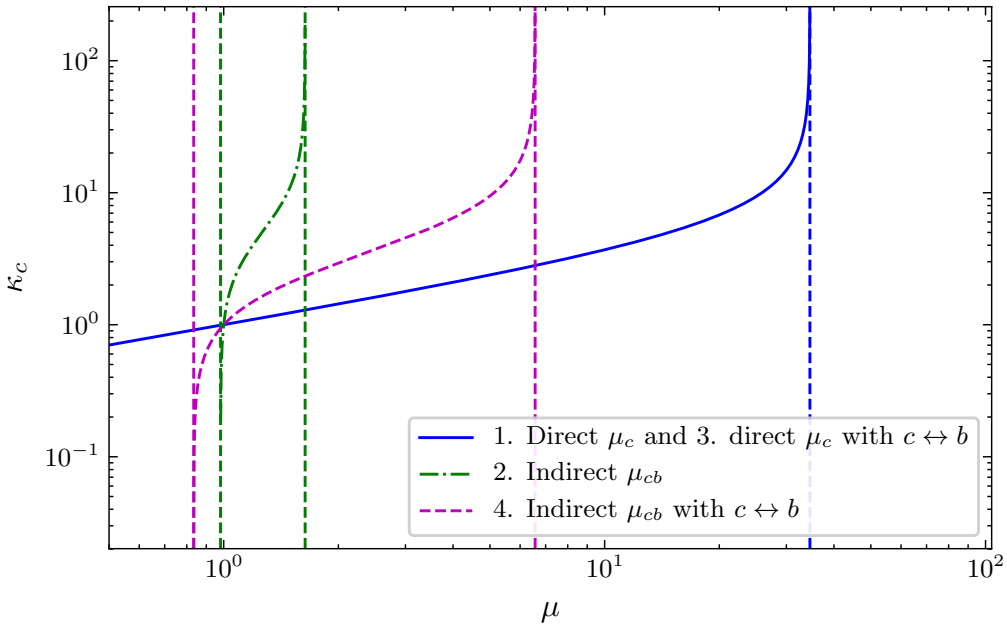


Figure 5.11: $\kappa(\mu)$ plotted in three instances. Blue: direct $H \to c\bar{c}$ defined in Eq. (5.5.1). Green: The indirect measurement defined in Eq. (5.5.4). Purple: The indirect measurement with charm-bottom discrimination defined in Eq. (5.5.6) evaluated with mean values of the efficiencies given in Tab. 5.3.

Expanding on ideas formulated for example in [85, 86], we also suggest an indirect measurement, in which $H \to c\bar{c}$ and $H \to b\bar{b}$ are combined to the signal class and

modified together with μ_{cb}. This extension transforms Eq. (5.5.1) into,

$$\mu_{cb} = \frac{\kappa_c^2 Br^{\rm SM}_{H\to c\bar{c}} + \kappa_b^2 Br^{\rm SM}_{H\to b\bar{b}}}{(Br^{\rm SM}_{H\to c\bar{c}} + Br^{\rm SM}_{H\to b\bar{b}})} \frac{1}{(1 + Br^{\rm SM}_{H\to b\bar{b}}(\kappa_b^2 - 1) + Br^{\rm SM}_{H\to c\bar{c}}(\kappa_c^2 - 1))} , \quad (5.5.4)$$

and the limitation to resolve κ_c becomes,

$$\frac{Br_{H\to b\bar{b}}}{(Br_{H\to b\bar{b}} + Br_{H\to c\bar{c}})} \frac{1}{1 - Br_{H\to c\bar{c}}} < \mu_{cb} < \frac{1}{Br_{H\to c\bar{c}} + Br_{H\to b\bar{b}}} . \quad (5.5.5)$$

If we further introduce $H \to c\bar{c}$ and $H \to b\bar{b}$ discrimination in the cut-flow with the $\mathrm{ML}_{b\leftrightarrow c}$ network Eq. (5.5.4) becomes,

$$\mu_{cb} = \frac{\epsilon_c \kappa_c^2 Br^{\rm SM}_{H\to c\bar{c}} + (1-\epsilon_b)\kappa_b^2 Br^{\rm SM}_{H\to b\bar{b}}}{(\epsilon_c Br^{\rm SM}_{H\to c\bar{c}} + (1-\epsilon_b) Br^{\rm SM}_{H\to b\bar{b}})} \frac{1}{(1 + Br^{\rm SM}_{H\to b\bar{b}}(\kappa_b^2 - 1) + Br^{\rm SM}_{H\to c\bar{c}}(\kappa_c^2 - 1))} . \quad (5.5.6)$$

the factors ϵ_c and $1 - \epsilon_b$ re-weight the expected event counts contributing to the signal strength based on the performance of the charm-bottom discriminator, $\mathrm{ML}_{b\leftrightarrow c}$. Now the limitation on κ_c is given by;

$$\frac{(1-\epsilon_b) Br_{H\to b\bar{b}}}{(1-\epsilon_b) Br_{H\to b\bar{b}} + \epsilon_c Br_{H\to c\bar{c}})} \frac{1}{1 - Br_{H\to c\bar{c}}} < \mu_{cb} < \frac{\epsilon_c}{(\epsilon_c Br_{H\to c\bar{c}} + (1-\epsilon_b) Br_{H\to b\bar{b}})} . \quad (5.5.7)$$

The limits of μ leading to a resolvable κ_c are determined for $\kappa_b = 1$. These expressions are illustrated in Fig. 5.11.

5.5.2 μ Results

In each channel we consider the primary contributions of uncertainties in the determination of μ. This can be done by fixing nuisance parameters in turn and varying each quantity by its pre-fit (initial uncertainties fed into likelihood) and post-fit (post fitted values from maximization of profiled likelihood) values.

In Fig. 5.12 we exhibit four classes of uncertainties,

1. Statistical,

2. Luminosity,

3. Systematics ("Sys"),

4. Monte Carlo ("MC").

Statistical uncertainties occur from the total counts in each bin in the likelihood function. The systematics from the 7-point envelope function in scale variations of f_F, f_R and lastly the Monte Carlo uncertainty from SHERPA. From this we can read off which backgrounds have the largest impact on the precise determination of μ: QCD, W+jets (W) production and Z+jets (Z) production processes for the 0L, 1L and 2L channels, respectively. It is worth stressing that measurements in the 2L channel, while being the most sensitive channel in this analysis, are dominated by the statistics from the limited cross-section of the signal process and lower luminosity.

We explored four ways for the μ extraction:

1. direct,
2. indirect,
3. direct with $b \leftrightarrow c$,
4. indirect with $b \leftrightarrow c$.

We summarise the interplay of the different analysis steps and the variation of μ_c and μ_{cb} in Tab. 5.5.

	cut-flow	$\text{ML}_{\text{booster}}$	$\text{ML}_{b\leftrightarrow c}$	κ dependence
1	✓	✓	X	$\mu_c(\kappa_c)$
2	✓	✓	X	$\mu_{cb}(\kappa_c, \kappa_b)$
3	✓	✓	✓	$\mu_c(\kappa_c)$
4	✓	✓	✓	$\mu_{cb}(\kappa_c, \kappa_b)$

Table 5.5: Summary of the cut-flows and neural networks architectures used in each of the methods and which of the κ_c, κ_b that are allowed to vary. Here $\text{ML}_{\text{booster}}$ refers to the ML "booster" network and $\text{ML}_{b\leftrightarrow c}$ the ML $b \leftrightarrow c$ discriminator network.

Fitting to distributions of various observables or to pairs of observables yields constraints for the the signal strengths μ_c. The following histograms over observables showed the best discriminating power:

- boosted subjet separation angle: $\mathcal{H}(\theta_{\hat{H},\hat{k}_{12}})$,
- fatjet mass: $\mathcal{H}(m_{\text{jet}})$,
- planar flow: $\mathcal{H}(P_f)$,
- 2-subjettiness: $\mathcal{H}(\tau_2)$,
- subjet energy fraction: $\mathcal{H}(z_1)$.

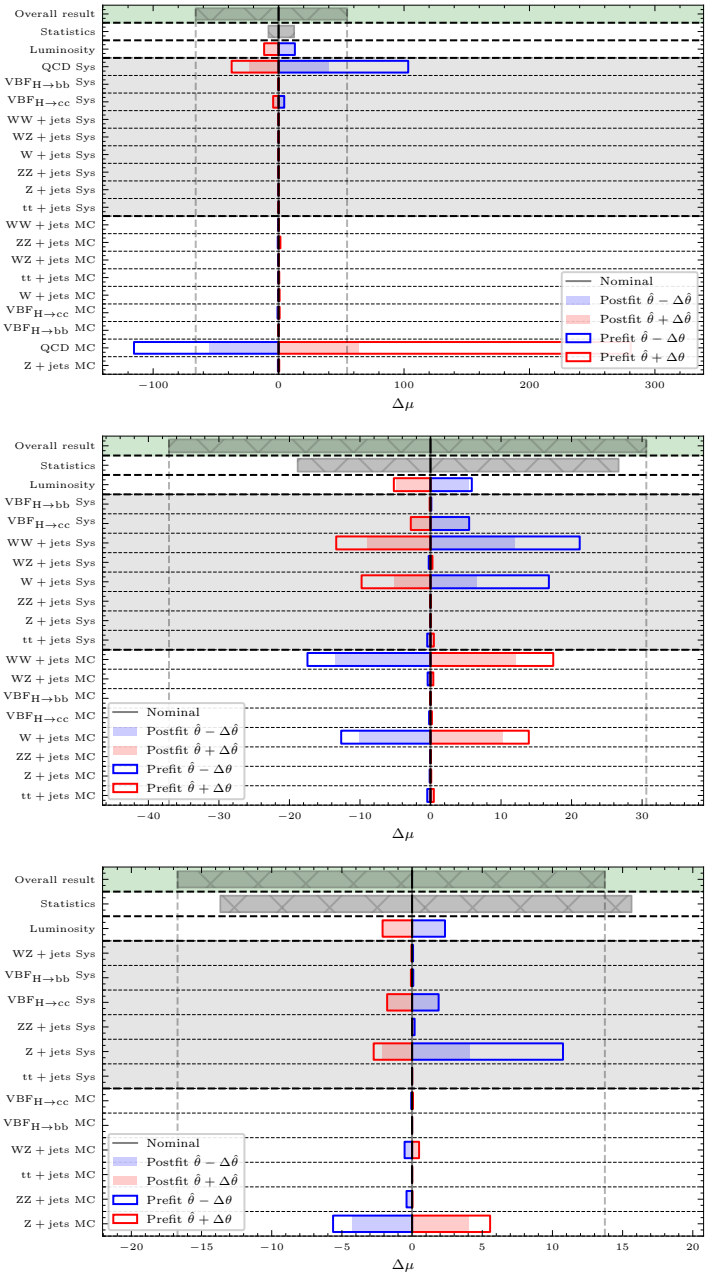
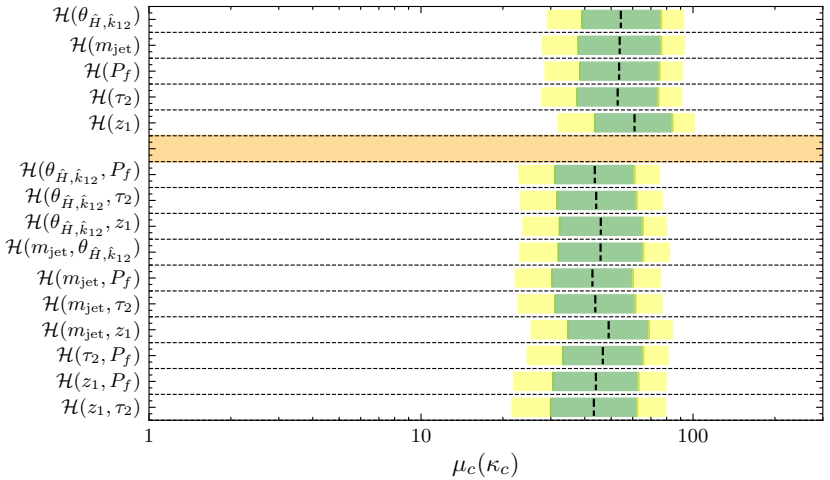
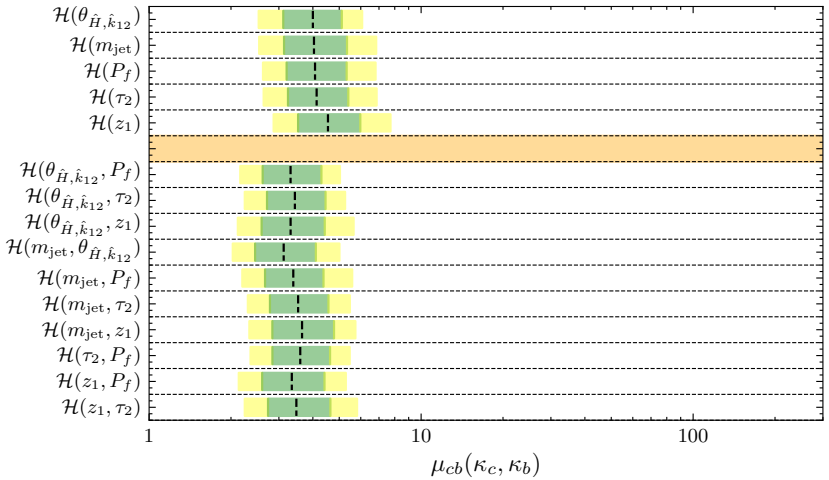


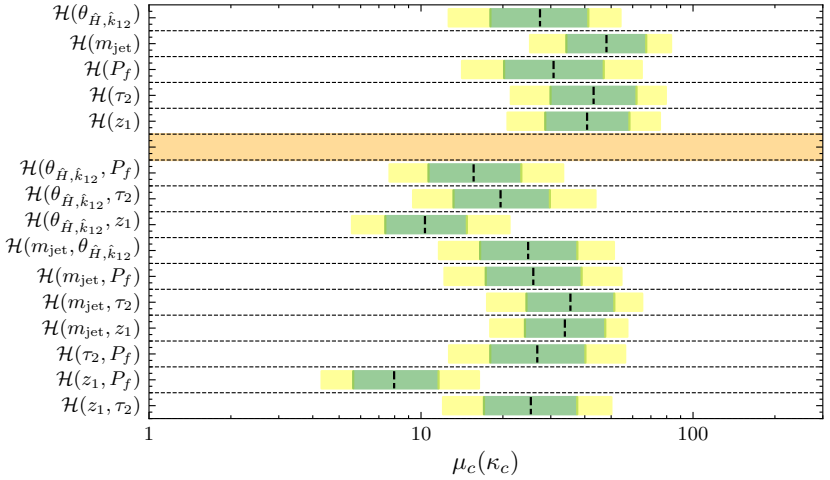
Figure 5.12: Uncertainty contributions to the 95% confidence limit of signal strength μ for $\mathcal{H}_{m_{\text{jet}}}$ for the jet mass distribution. Uncertainties are shown for total statistics, luminosity, systematic uncertainty and Monte Carlo simulation uncertainty over all signal and background classes for each channel and compared with the total uncertainty. Pre-fit uncertainties shown in the outlined bars and post-fit uncertainties by the filled in bars. Top: 0L. Centre: 1L. Bottom: 2L.



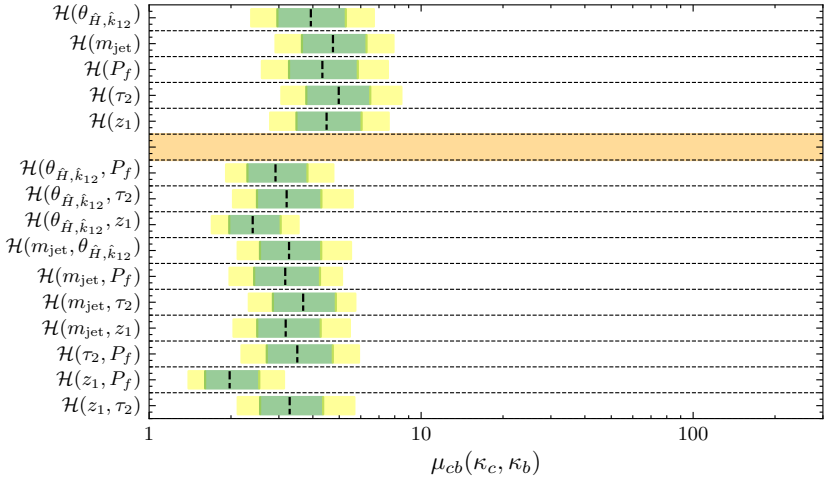
1. Direct



2. Indirect

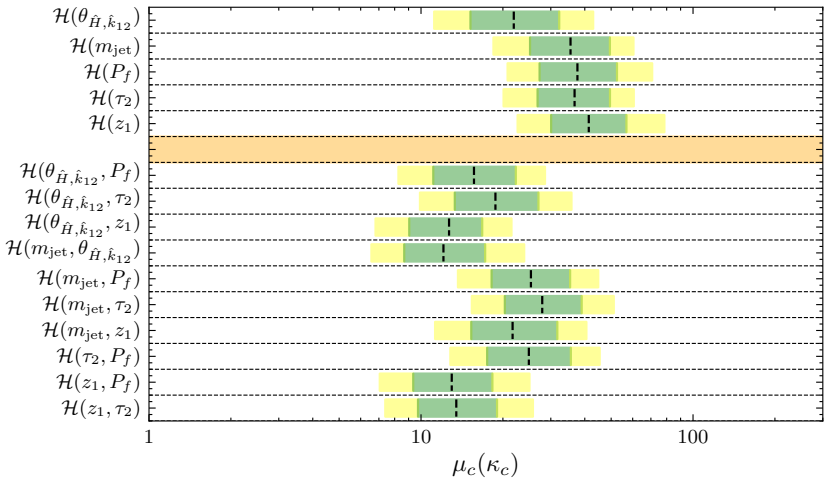


3. Direct with $b \leftrightarrow c$ discrimination

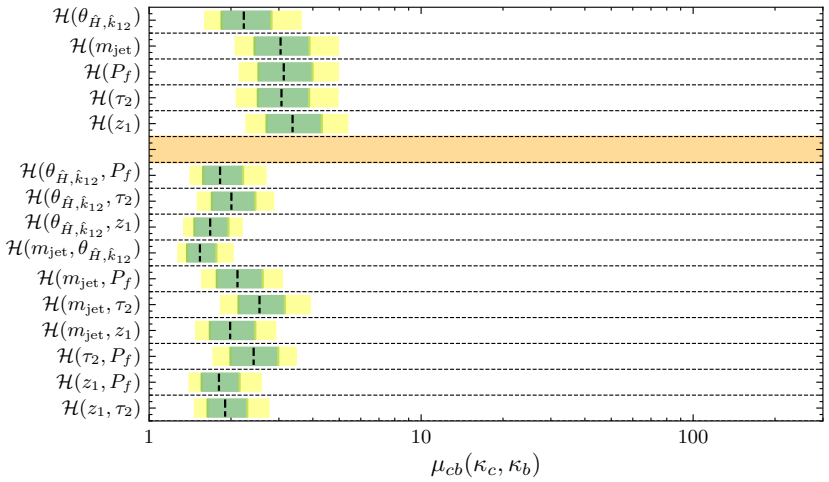


4. Indirect with $b \leftrightarrow c$ discrimination

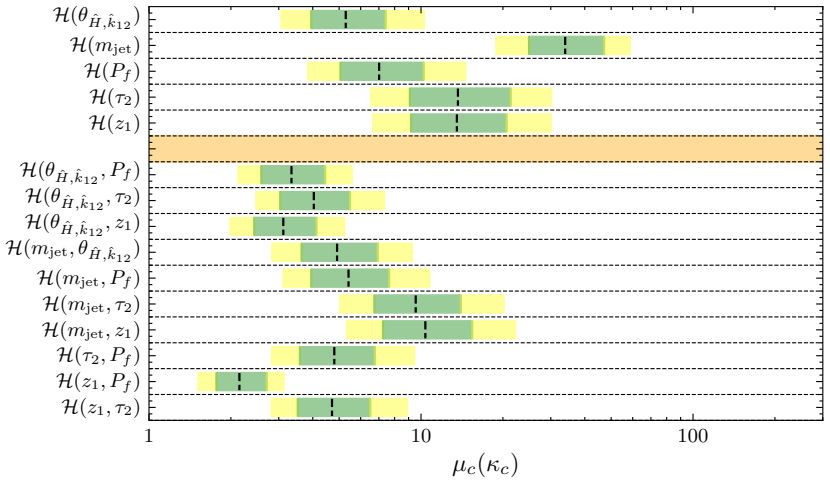
Figure 5.13: Comparison over binned likelihood distributions showing the 95% confidence limit on the signal strength μ at $\int \mathcal{L}dt = 150\text{fb}^{-1}$ and the green and yellow colours bars showing its 1σ and 2σ errors.



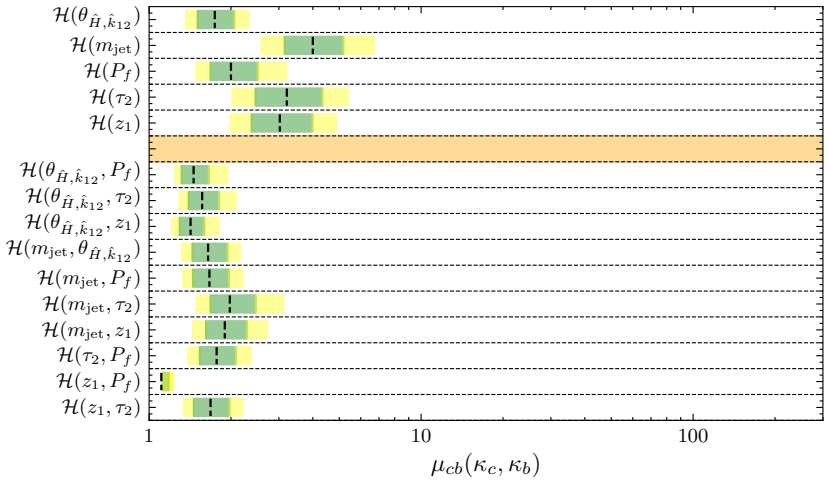
1. Direct



2. Indirect



3. Direct with $b \leftrightarrow c$ discrimination



4. Indirect with $b \leftrightarrow c$ discrimination

Figure 5.14: Comparison over binned likelihood distributions showing the 95% confidence limit on the signal strength μ at $\int \mathcal{L}dt = 3000\text{fb}^{-1}$ and the green and yellow colours bars showing its 1σ and 2σ errors.

We present results for the four methods over all distributions for an integrated luminosity of $150\,\text{fb}^{-1}$ in Fig. 5.13 and for $3\,\text{ab}^{-1}$ in Fig. 5.14. The figures exhibit the 95% confidence limits of all considered 1-dimensional fits and two-dimensional fits with the 1σ and 2σ uncertainty bands.

Search		μ w/ $\mathcal{H}_{m_\text{jet}}$	μ w/ $\mathcal{H}_\text{best}$	$\mathcal{H}_\text{best}$
$\int \mathcal{L}dt = 150\,\text{fb}^{-1}$				
1.	Direct, μ_c	$53.7^{+22.7}_{-15.9}$	$42.7^{+17.4}_{-12.4}$	$\mathcal{H}(m_\text{jet}, P_f)$
2.	Indirect, μ_{cb}	$4.0^{+1.3}_{-0.9}$	$3.1^{+1.0}_{-0.7}$	$\mathcal{H}(m_\text{jet}, \theta_{\hat{H},\hat{k}_{12}})$
3.	Direct, μ_c with $b \leftrightarrow c$ discrimination	$48.1^{+19.2}_{-13.8}$	$8.0^{+3.6}_{-2.3}$	$\mathcal{H}(z_1, P_f)$
4.	Indirect, μ_{cb} with $b \leftrightarrow c$ discrimination	$4.7^{+1.6}_{-1.1}$	$2.0^{+0.6}_{-0.4}$	$\mathcal{H}(z_1, P_f)$
$\int \mathcal{L}dt = 3000\,\text{fb}^{-1}$				
1.	Direct, μ_c	$35.5^{+14.0}_{-10.3}$	$12.1^{+5.1}_{-3.4}$	$\mathcal{H}(m_\text{jet}, P_f)$
2.	Indirect, μ_{cb}	$3.0^{+0.8}_{-0.6}$	$1.5^{+0.2}_{-0.2}$	$\mathcal{H}(m_\text{jet}, \theta_{\hat{H},\hat{k}_{12}})$
3.	Direct, μ_c with $b \leftrightarrow c$ discrimination	$33.9^{+13.2}_{-8.9}$	$2.1^{+0.6}_{-0.4}$	$\mathcal{H}(z_1, P_f)$
4.	Indirect, μ_{cb} with $b \leftrightarrow c$ discrimination	$4.0^{+1.2}_{-0.8}$	$1.1^{+0.1}_{-0.1}$	$\mathcal{H}(z_1, P_f)$

Table 5.6: Combined μ bounds at 95% confidence for four methods for the $\mathcal{H}_{m_\text{jet}}$ and $\mathcal{H}_\text{best}$ distributions. We compare results obtained from the fatjet distribution ($\mathcal{H}_{m_\text{jet}}$) – the method of choice for the experimental analyses so far – with results we obtained from a best combination of two observables, $\mathcal{H}_\text{best}$ and indicate them in the last column.

One way in which the impact of dominant statistical uncertainties can be counteracted is by performing a two-dimensional fit and constraining the sum of bins on each axis to one another and therefore their uncertainty to one another. We summarize the μ limits we obtain in Tab. 5.6 for the standard choice $\mathcal{H}(m_\text{jet})$, used by the experimental analyses so far and our most powerful combination of distributions, $\mathcal{H}_\text{best}$. Due to the small branching fraction of $H \to c\bar{c}$ we have a very low signal count at low luminosity. Moving forward into the high luminosity phase of the LHC with around $3000\,\text{fb}^{-1}$ we will be less limited by statistics. This impacts on the

efficiency of the ML "booster" cuts, which greatly improve our confidence limit by a factor of two over the initial cuts at $150\,\text{fb}^{-1}$.

The two-dimensional fitting technique leads to an improvement in the obtained confidence limits by (on average) a factor of 2 over their one-dimensional counterparts. We also see that while distributions involving m_{jet} provide good fits, other choices of observable work just as well or even better hinting at possible new avenues of exploration. The ML $b \leftrightarrow c$ discriminator network provides an improvement to the value of the confidence limit of a factor 2 in the direct case but only 1.1 in the indirect case. Our best fit result is $\kappa_c \leq 3.18^{+0.94}_{-0.60}$ ($\mu_c \leq 8.0^{+3.6}_{-2.3}$) at $150\,\text{fb}^{-1}$ at the 95% confidence limit. This result is compatible to SM within 4.0 standard deviations and is competitive with current ATLAS and CMS values of $\kappa_c \leq 8.5^{+3.9}_{-3.9}$ [111] and $\kappa_c \leq 3.4$ [91] determined at $139\,\text{fb}^{-1}$ and $138\,\text{fb}^{-1}$ respectively. These results may have scope for enhancement by considering a wider range of features and multi-dimensional fitting rather than the using the "standard" $\mathcal{H}(m_{\text{jet}})$ choice as we have demonstrated in these findings.

Moving into the high luminosity regime we see again an enhanced benefit from two-dimensional fits by a factor of $\sim$2 over one-dimensional fits. At $3\,\text{ab}^{-1}$ our best fit μ values tighten and the limits become resolvable under the κ framework. The direct measurement provides the best expected upper-bound of $\kappa_c \leq 1.47^{+0.21}_{-0.16}$ which is again competitive against the ATLAS projection at $3\,\text{ab}^{-1}$ for the 95% confidence limit of $\kappa_c \leq 3.0$ [111].

5.6 Conclusions

We studied prospects for a determination of the charm Yukawa coupling and its constraints with present and future LHC data. We considered the production of the Higgs boson in Higgs-Strahlung and weak boson fusion processes, leading to final states with 0, 1, or 2 leptons and the subsequent decay of the Higgs boson into a fatjet. We augmented a simple cut-based strategy with a multivariate "booster" step

and showed that this enhances the sensitivity of the analysis. We also investigated the impact of a neural network based discriminator for fatjets containing only light partons, charm or bottom quarks and found a measurable impact. As a by-product, we suggested an indirect measurement strategy where the branching ratio of a Higgs boson into heavy quarks – charm or bottom – is used in conjunction with the known value of the bottom Yukawa coupling to infer the charm Yukawa coupling. The signal strength in any case is extracted from fits to observable distributions and it is shown that the fatjet mass is not necessarily the best-suited observable. Further it is found that two-dimensional fits can further boost the sensitivity by about a factor of two to four compared to fits to a single observable, motivating further investigations.

Part III

Industrial Placements

Chapter 6

Anomaly Detection at Hartlepool Power Station

6.1 Introduction

Hartlepool Power Station is located in the North East of England and is capable of providing enough power for two million homes. It consists of two advanced gas-cool reactors (AGR), which have been operational since 1983 [112] and is one of four sites using AGR style reactors in the UK. There are a vast array of sensors and monitoring techniques of the reactor and its health which ensure the safety standards at Hartlepool are very high. There exist multiple fail-safes in its operation and the sensors have conservative operational parameters which are closely monitored at all times to ensure the reactor remains safe [113].

We investigate a set of tools that have the potential to better inform engineers of changes in the operational parameters of a nuclear power plant. These changes could be indicative of unexpected load drops or be indications of a developing fault in the essential sensors and instrumentation. An unexpected load drop is any event which leads to a drop in reactor power output that could be indicative of a fault in the reactor systems. In these drops one of two things occur: the reactor automatically

trips which brings the nuclear reaction to a halt safely or the engineers trip the reactor manually to mitigate any potential risk. There are other reasons that the reactor would be brought offline than unexpected load drops: routine refuelling or maintenance – data in these regions we consider to be *healthy*, the reactor is behaving as we nominally expect. We are interested in comparing these *healthy* regions with those regions prior to a known trip which we will refer to as *monitor* regions. The goal is to build a framework of tools packaged into a demonstration dashboard in which an engineer could monitor a live feed of the variables measured in systems of the reactor. This would better inform them of any potential changes in the structure of the data that could be indications of faults developing and so prompt them to manually check for issues. In this work we outline the methods and details of the mathematical tools used and their successes. Then in the Appendix B we provide a user guide for the dashboard.

The data is segmented into different regimes based on the reactor operation at that time. In this section graphs will contain shaded portions indicating the different regions:

- *monitor*: Regions prior to a known unplanned trip highlighted in green.
- *healthy*: Regions in which the reactor is operating nominally at full capacity for over one week prior to refueling and they are indicated in blue.
- down: Regions where the reactor has $\leq 40\%$ power output in red.

Fig. 6.1 left shows an example *healthy* region which shows the power output slowly being manually lowered over the course of a $\sim$4+ hour period until the reactor is deliberately tripped at around 40% capacity. This is in contrast to the right in which the reactor automatically is tripped from around 100% capacity almost instantaneously. These typical regimes of operation can be used as a template to manually collect a dataset of these regions prior to these load drops which we can then confirm against the reported trip logs.

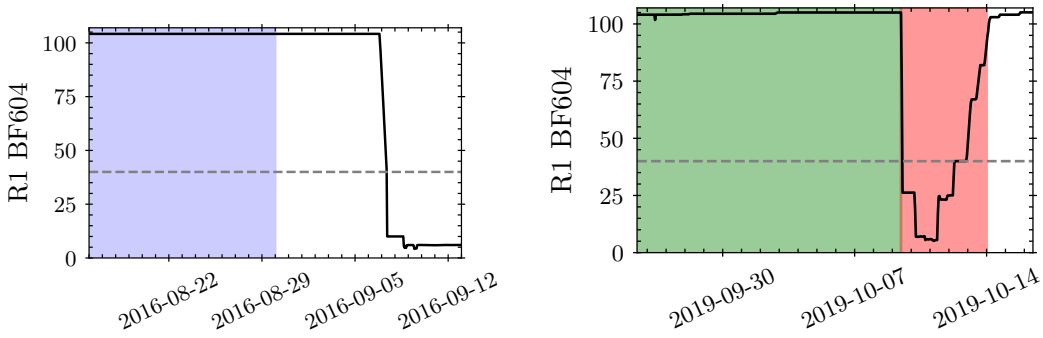


Figure 6.1: Percentage of maximum reactor power output "R1 BF604".
Left: a *healthy* refuelling region. Right: a trip *monitor* region.
The highlighted colours indicate data regimes used in training predictive models. Blue: A *healthy* region. Green: A *monitor* region. Red: Period of reactor downtime.

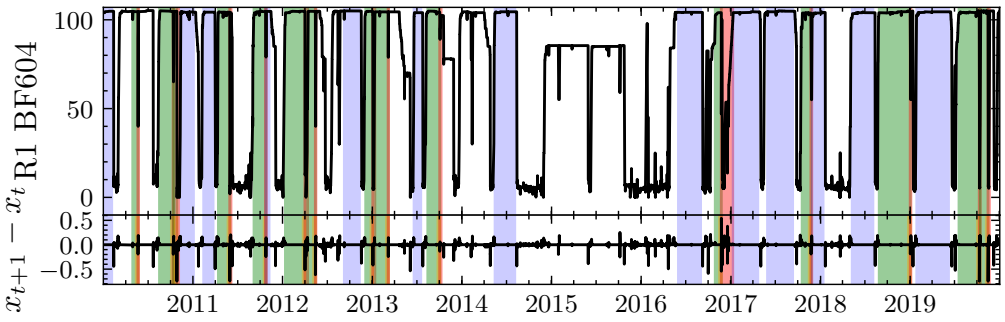


Figure 6.2: Top: Reactor 1 power output data over the full data range, 2010 – 2020. Bottom: Difference between sequential hourly measurements.
The highlighted colours indicate data regimes used in training predictive models. Blue: A *healthy* region. Green: A *monitor* region. Red: Period of reactor downtime.

In Fig. 6.2 we see a set of highlighted regions indicating the data we will use for developing our analysis tool chain, all other data are discarded. This discarded data includes reactor downtime or regions which have a duration of less than 1 week. The data kept are indicated by the blue and green regions which are those known to be *healthy*, or leading up to a trip and are also at least 1 week in duration. In 2015 Reactor 1 was operating at a reduced capacity so as to limit boiler temperatures

until certain modifications to the reactor had been conducted. This can be seen clearly in Fig. 6.2 between 2015—2016. As this region is not representative of typical reactor operation we discard the data from this time. In particular we have 15 and 13 trip events in Reactor 1 (R1) and Reactor 2 (R2), respectively and there are also 13 and 12 *healthy* refuelling regions in R1 and R2, respectively these *healthy* data are truncated 1 week prior to the refuelling as in this week the reactor is prepared for going offline; this preparation may alter the nominal behaviour of many of the measured variables so we discard data from this region.

The data presented in this section graphically are collected from R1 only and from two reactor systems specifically, the condenser and vacuum pumps. Later in Section 6.5 we will also consider the Channel Gas Outlet (CGO) dataset monitoring the fuel rod channels. Each reactor is functionally identical to one another, so for now we will examine the data from one. We can treat each reactor as independent identical systems so any R1 trained model would be functional for R2. A subset of possible measurements from each system is considered, the condenser data has 11 features and the vacuum subsystem consists of 28 features which include measurements such as feed and outlet water temperatures, pressure and dissolved chemical concentrations. These features are good indicators of system health and should be stable in nominal operation. Our datasets are pulled from 10 years of historical data in the years between 2010 and 2020 and we sample the data over 1 hour windows in the interest of keeping computational cost down during the data exploration. We calculate the mean and standard deviation over each hour interval as an estimate of the value and error of that data point during that period. In fact many of the measurements are sampled as frequently as every few seconds, giving scope for more granular or coarse monitoring. An example feature is plotted in Fig. 6.3 for the R1 condenser inlet temperature and as shown by the 5σ error bars the data is relatively stable over each hour window except in some instances, indicated by the large highlighted error bars. In some cases measurements are “missing” or “nan” values (those which are faulty), in which case these values are filled in with quadratic interpolation to give a good

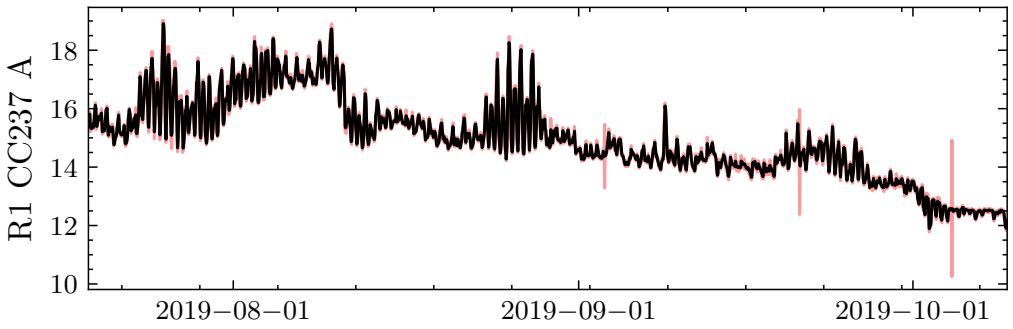


Figure 6.3: "R1 CC237 A", Condenser inlet temperature over a portion of a *monitor* region. The red highlighted region indicate the 5σ error bars.

estimator for the value. However, it is noted that these points are almost entirely contained within reactor downtimes and usually isolated from one another so they can be interpolated from their neighbours. Fig. 6.3 also gives a good indication of the challenge faced here, there are varied levels of periodicity in the data, high levels of noise and as will be shown there are many dependencies between features and external parameters. We will tackle these issues with a number of standard time series analysis tools and a novel approach which bootstraps periodicity into our analysis.

In the interest of presentation and brevity, we will follow the example *monitor* region 19/7/2019 - 9/10/2019 (Fig. 6.3) throughout the presentation of the tools and methodologies. Further any plots in which we consider an analysis technique involving just one feature we will use "R1 CC237 A" which is a measure of a condenser inlet temperature.

6.2 Tools and Techniques

In order to best tackle the task of anomaly detection we need to get around a number of challenges typically associated with any time series analysis involving time dependent datasets. Typical approaches take the time series and perform a number of transformations to produce a "stationary time series" [114]. This allows

for more stable forecasting and a more detailed understanding of the time series. A stationary time series is one which is strictly independent of time for example a random walk, or white noise. Producing a stationary time series from a non-stationary time series can be done by methods of differencing, removing trends and seasonality or parametrisation. The reactor data has seasonality predominantly on yearly, weekly and daily scales. Methods of differencing are powerful for long and complete datasets as is fitting seasonality and trends, however we would like to be able to examine the data over shorter timescales. By considering the data over short time scales we avoid preprocessing all of the historical data, potentially many times which quickly becomes computationally expensive with higher dimensional data. Here the approach that we focus on is parameterisation. We examine the frequency spectrum of the data provided by Fourier transforms over a weekly rolling windows fitting to a simple power law model – making seasonality in the data inherent. All analyses are implemented in PYTHON using the standard libraries that can be installed with PIP, the PYTHON package manager.

6.2.1 Frequency Analysis

The frequency spectrum of each feature is analysed using the discrete Fourier transforms (DFT) which extracts a frequency decomposition composing the original data into frequency components f and their amplitudes A. The discrete nature of the data and its length constrains the possible frequencies in the range $1/2T \leq f \leq 1/2\Delta T$ with $\Delta T = 1$ hour and $T = 10$ years, in the most extreme range. DFT is defined for a set of N data points x_n measured at time $t = n\Delta T$ and is given by the following equation,

$$X_k = \sum_{n=0}^{N-1} x_n \exp\left(\frac{-i2\pi}{N} nk\right) \quad \text{for} \quad k = 0, 1, \ldots, N-1 \,, \tag{6.2.1}$$

where X_k are complex valued Fourier transformed variables of the data x_n and $k = 0, 1, \ldots, N-1$ are directly mapped to evenly spaced frequencies defined by the sampling rate, $f = 0, 1/2T, \ldots, 1/2\Delta T$. The amplitude A_k and phase ϕ_k are defined

for a given frequency:

$$A_k = \frac{1}{N}|X_k| \,, \tag{6.2.2}$$

$$\phi_k = \arctan\left(\frac{\mathrm{Im}(X_k)}{\mathrm{Re}(X_k)}\right) . \tag{6.2.3}$$

The data can now be expressed in the form,

$$x_n = \sum_{k=0}^{N-1} A_k \cos(2\pi f_k t + \phi_k) \,. \tag{6.2.4}$$

The Fourier transformed data are fit to the following power spectrum model which depends on $A|_{f=1}$ and α using χ^2 fitting [115],

$$A = A|_{f=1} f^{\alpha} \,, \tag{6.2.5}$$

where $A|_{f=1}$ is the amplitude at $f = 1$ Hz and α is some exponent characterising the frequency power spectrum. This equation defines a power law frequency spectrum, there are two special cases:

- $\alpha = -1$: White Noise. Uniform energy density per frequency interval,
- $\alpha = -2$: Pink Noise. Uniform energy density per octave.

Pink noise power spectra are remarkably common in nature [116] and provides a way to make seasonality in our data inherent as it is found that all of the features fit this frequency power law nicely. This power spectrum equation can be fit over a rolling window across the data, thus we can extract a new time dependent feature $\alpha(t)$ which will remain constant if the frequency spectra remains constant. The standard error of the data is used to inform the χ^2 fit [115],

$$\alpha_{x_t} = \frac{\sigma(x_t)}{\sqrt{N_{\text{window}}}} \,. \tag{6.2.6}$$

This error can then be associated with the amplitudes A_k, this will then reduce the affects of low amplitude oscillations on the fitting which could be associated with noise not a physical process where the magnitude of the uncertainty is the data is comparable to the magnitude of the noise.

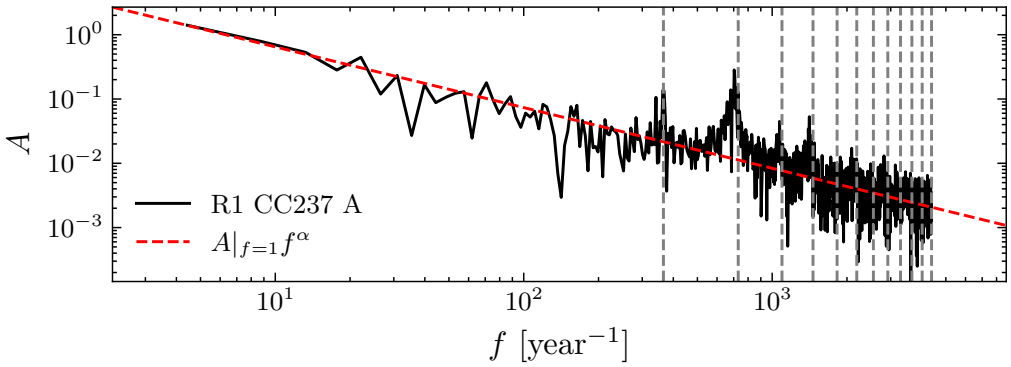


Figure 6.4: The frequency spectrum of "R1 CC237 A" and its best fit line for the power spectrum. The vertical lines indicate key harmonics of multiples of daily oscillations $f_h = N\text{day}^{-1}$.

Fig. 6.4 shows an example pink noise fit on one feature "R1 CC237 A" over an entire *monitor* region. We see strong peaks at multiples $f = N\text{day}^{-1}$, indicating harmonic oscillations over fractions of the day which is seen in the raw feature plot Fig. 6.3. We are interested in how the exponent of this fit, α changes over time. Changes in the frequency structure of the data will be reflected in changes in α and its quality of fit therefore the oscillatory structure in the data is inherent. However it is important to note the χ^2 fit is only sensitive to the following frequencies $f_{\text{min}} = 1/2T_{\text{window}}$. A sufficiently large window should be chosen so that larger frequencies we to understand are well represented in the model fit.

6.2.2 Motivating Frequency Decomposition

In this section, we consider two methodologies of unsupervised learning that we introduced in Chapter 2 PCA and BOCPD and why these methods require stationarity, that is the underlying distributions generating the data are static and de-trended. As outlined in the introductory chapters to this thesis, PCA and BOCPD tools provide a method of probabilistic interpretation by the T^2-Hotelling and Q statistics and $\boldsymbol{L}(l,t)$ matrix. In particular the easy interpretation is particularly important in this context, while NNs may be quicker to set up and deploy they can often be

considered black-boxes in which we gain little or no insight into the underlying data. This is something that could be considered problematic at a nuclear power station.

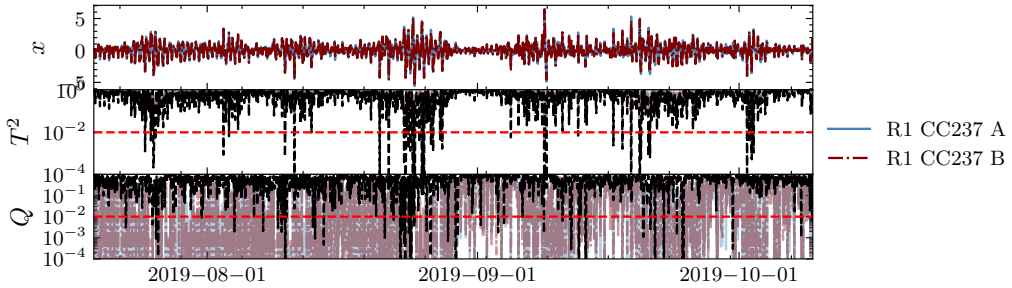


Figure 6.5: An example trip *monitor* region for the feature space "R1 CC237 A" and "R1 CC237 B". Top: Normalised variables, Centre: p-values of the T^2 statistic, Bottom: p-values of the Q statistic. Red line indicates the confidence limit, $\alpha_{\text{crit}} = 0.01$. Black line indicates the overall value of T^2 or Q and the colour lines are the marginalised values over the features.

Fig. 6.5 shows an example *monitor* region over time for a two variables and its T^2 and Q statistics using a "rolling PCA" model which will be explained in detail in Section 6.3.2 using the raw data in which there is no attempt at de-trending. There are issues with using PCA over the full duration or a very large window, the raw data is non-stationary over long time intervals (i.e. daily, yearly effects caused by weather and power demands). PCA is not appropriate with non-stationary data, a problem which we will use frequency analysis to address. Further to this, a long window duration which contains anomalous data will have its model behaviour altered and performance impacted by said anomalous points. This non-stationarity can be seen in the periodic oscillations the top subplot in Fig. 6.5 and the periodic drop in T^2 and Q p-values.

Next if we instead apply BOCPD (see Fig. 6.6) we see the raw data generates a large collection of CPs because we have failed to de-trend our data. Regardless, it is however possible to see from the lighter portions of the heat map a saw-tooth pattern over longer periods of time which indicates longer period changes to the structure of the time series. Another step we can take to aid the BOCPD in dealing with slow

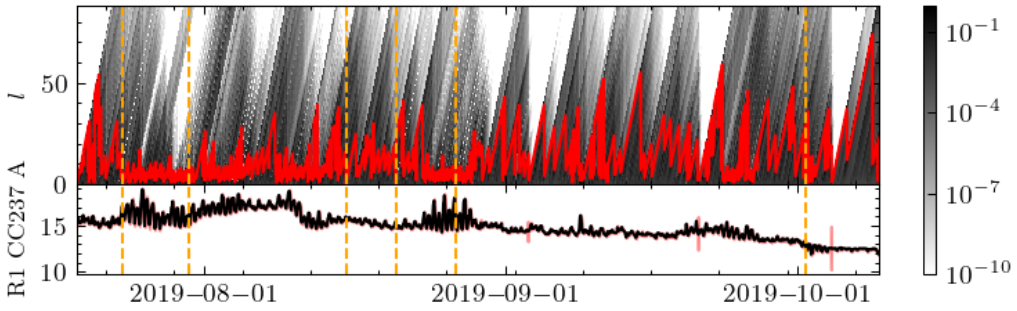


Figure 6.6: BOCPD for an example trip *monitor* region for feature "R1 CC237 A". Dashed orange lines indicate CPs, when $l_{max} = 0$.
Top: The colour map of matrix $\boldsymbol{L}(t, r)$ with red indicated l_{max} where the probability of a new sequence is largest. Bottom: The example feature "R1 CC237 A" before a trip.

continuous trends is to take a first order differences. That is, we take the difference between neighbouring time-series data points which has the effect of removing linear trends. The tolerance of the BOCPD for noisy data is captured by the values of σ_t^l, to make the analysis more forgiving to larger deviations from the mean we can scale the σ by our desired confidence. We multiply the standard deviation σ of the data in a sample sequence by a scaling factor of 5 to give a wider tolerance. In the following section we combine these tools and techniques and act them on our data.

6.3 Methodology

In the previous sections, the main tools have been explored with a set of examples and plots and we have justified the requirement of de-trending. Now we combine the tools into a set of complementary anomaly detection frameworks using the entirety of the data and features together. A final preprocessing step that is found to be particularly powerful is taking a first order differences which is performed on both BOCPD methods and the PCA data before being standardised. This helps to deal with slow moving long term linear trends.

6.3.1 α-Analysis

We can use Eq. (6.2.5) to parameterise our data using α exponent over a rolling time window, any changes in frequency space of the data will be reflected in changes in $\alpha(t)$. The time scale for the rolling window of one week is chosen, this gives a reasonable number of data points to perform our fit to the power spectrum equation and enough points to "thermalise" our models making them less susceptible to noise in the data over small time scales while providing prompt forecasting.

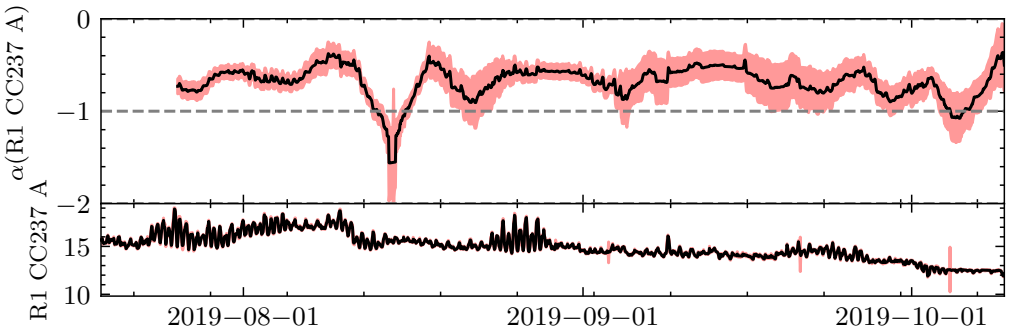


Figure 6.7: Example $\alpha(t)$ fit for an example trip *monitor* region for α analysed "R1 CC237 A".
Top: $\alpha(t)$ fit over "R1 CC237 A". Bottom: "R1 CC237 A" before a trip. Highlighted regions indicate the 5σ error bounds.

Using this rolling window method we do however sacrifice the first window duration for forecasting to determine $\alpha(t)$ over the first window, this is why the top portion of Fig. 6.7 starts later than the raw feature in Fig. 6.7. Similarly, in BOCPD the first window is required to build the priors in the model.

The BOCPD σ values are determined from the χ^2 fit parameter errors and scaled to 5σ confidence to give a wider tolerance for long sequences with noisy data. Fig. 6.6 compared with Fig. 6.8 shows better stability in extracted CPs compared with the original raw feature. Looking for commonality in CPs between related features will provide insight into specific areas of the systems which may have altered. This alone is not enough to say anything substantial about the measurements; we look to PCA to give indications of how anomalous the features may be. However, in the case of

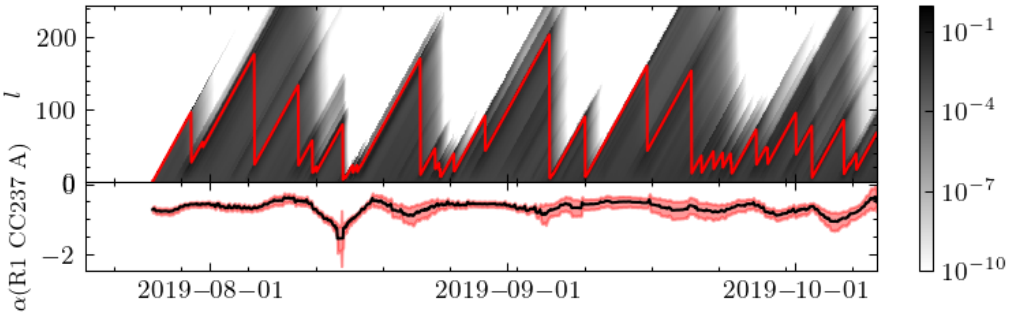


Figure 6.8: BOCPD for an example trip *monitor* region for α analysed "R1 CC237 A".
Top: The colour map of matrix $\boldsymbol{L}(t, r)$ with red indicated l_{max} where the probability of a new sequence is largest. Bottom: The example $\alpha(t)$ of "R1 CC237 A" before a trip.

"R1 CC237 A" the strong change points occur when changes in the periodic structure of the data change.

6.3.2 Rolling Window PCA

We implement a PCA method similar to those implemented in [16–18] which informs our certainty in changes in the operation of R1 and R2. Once the model is fit, we calculate the T^2 and Q statistic live which forms a "trigger" for regions which may require the attention of an engineer. Our trigger is,

$$\alpha_{T^2} < \alpha_{\text{crit}} \quad \text{and} \quad \alpha_Q < \alpha_{\text{crit}} \,. \tag{6.3.1}$$

In this work, unless specified otherwise, we take $\alpha_{\text{crit}} = 0.01$ which corresponds to a 99% confidence limit that the data is anomalous. In these conditions we catch the extremal data with respect to the PCA models which are fit over a rolling window which error checks the subsequent new data point and its T^2 and Q statistic before refitting in the following window. Further to this trigger we can require a minimum number of consecutive data points in which these criteria are met i.e. multiple anomalous data points in a row. Any regions with at least one hour of triggered data points are highlighted in orange. A powerful feature of this methodology is

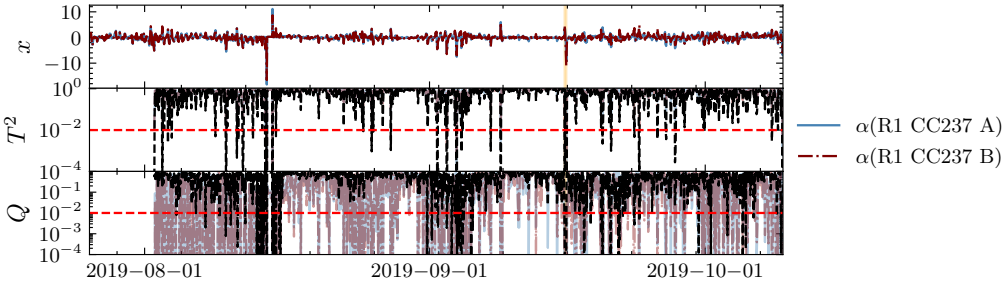


Figure 6.9: An example trip *monitor* region for the α processed feature space "R1 CC237 A" and "R1 CC237 B".
Top: Normalised variables, Centre: p-values of the T^2 statistic, Bottom: p-values of the Q statistic. Red line indicates the confidence limit, $\alpha_{\text{crit}} = 0.01$. Black line indicates the overall value of T^2 or Q and the colour lines are the marginalised values over the features.

the potential to marginalise T^2 and Q over individual input features. This can be done by calculating the T^2 and Q statistic for new data point $\boldsymbol{x}$ and replacing all feature values with their mean in the current window $\mu(x_f(t - T_{\text{window}} : t))$ except the chosen feature x_f. This provides a deeper insight into which feature is dominant in producing the anomaly or if many or all of the features are correlated. In Fig. 6.9 we see highlighted regions (14th August 2019, 2nd September 2019, 15th September 2019) which are not successfully highlighted by Fig. 6.5 using only one feature input alone. However looking at Fig. 6.8 we see clear indications of CPs from l_{max} dropping in value significantly at these dates except in 2nd September instance in which the raw data had a large error bar (see Fig. 6.3) thus BOCPD has not considered this a CP. We see that between these two methods we have extracted the dates where the distributions have changed or if there is an impulse spike in the data or its error. The PCA and BOCPD models acting on the α - analysis data have picked these out and highlighted them for the user. However it is important to note this does not necessarily indicate a fault, that is for the engineer to confirm. It is also useful to note that the power of the PCA analysis improves with more features, considering one or two features makes the analysis more sensitive to statistical noise especially

in features which have large oscillations.

BOCPD is implemented in the dashboard but we will discuss it no further and focus on the PCA methodology and results as the time constraints of the placement took us in that direction. However, the BOCPD provides a promising avenue for monitoring tools in particular it could be expanded into higher dimensions to study correlated features.

6.3.3 Improvements to Rolling Window PCA

As noted in the previous section, the PCA detection is only sensitive to changes if the window prior contains mostly nominal behaviour. Therefore if we detect an anomaly that satisfies a very strict α_{crit} threshold the next window is less sensitive to comparable anomalies because the previous data point which was anomalous has "inflated" the standard deviation of the windowed data. To get around this we implement a rolling average flag rate of anomalous data $\bar{\beta}$ where we instead use a different α_{crit} which has been carefully chosen by generating ROC curves (cf. Section 2.4). The rolling average flag rate is calculated,

$$\bar{\beta}(t_1) = \frac{\sum_{t_i=t_0}^{t_1} \beta\Big(\alpha_{T^2}(t_i), \alpha_Q(t_i)\Big) W\Big(t_1 - t_i\Big)}{t_1 - t_0} \,, \tag{6.3.2}$$

where, $\beta(\alpha_1, \alpha_2)$ is a Boolean value such that,

$$\beta(x, y) = \begin{cases} 1 & \text{if } x < \alpha_{\text{crit}} \text{ and } y < \alpha_{\text{crit}} \\ 0 & \text{if } x \geq \alpha_{\text{crit}} \text{ and } y \geq \alpha_{\text{crit}} \end{cases} . \tag{6.3.3}$$

We also have $W(t_1 - t_i)$ which is a window smoothing function, providing a greater weighting on nearer time points if required. For the rest of the work we define $W(t_1 - t_i) = 1$ therefore Eq. (6.3.2) transforms to a unweighted rolling mean. We can now demand that this rolling rate exceeds a new threshold,

$$\bar{\beta} > \bar{\beta}_{\text{crit}} \,, \tag{6.3.4}$$

which tells us that the rate of anomalous data points in the time series has reached some critical rate in which we can consider the previous rolling window of data as anomalous and requiring investigation.

6.3.4 Optimization

We can optimise our model hyper parameters such as α_{crit} using the metrics accuracy, precision, recall or F1 score depending on what section of the confusion matrix we care about most (cf. Section 2.4). F1 score is the most useful as we want to ensure we have a balance between the precision and recall which come at the expense of one another. We will use ROC curves to inform our choices of the hyper parameters along with the confusion matrices. Our optimization for rolling window rate threshold is outlined in Alg. 2.

Alg. 2: Algorithm to pick the optimum α_{crit} and $\bar{\beta}_{\text{crit}}$.

Input **Data:** $\alpha_{\text{crit},i} = 0$

for $\alpha_{crit,i} \in [10^{-8}, 0.9]$ **do**

  Determine anomalous points with T^2 and Q thresholds set at $\alpha_{\text{crit},i}$

  Produce ROC curves with $\bar{\beta}$ forming the decision boundary

  Calculate AUC

end

Set α_{crit} from $\alpha_{\text{crit},i}$ with greatest AUC ROC curve

Pick $\bar{\beta}_{\text{crit}}$ optimising for the chosen metric and α_{crit}

The hyper-parameters tuning to pick α_{crit} and $\bar{\beta}_{\text{crit}}$ is executed on a subset of the labeled data using the typical reactor profiles outlined in the introduction.

6.4 Results

It is difficult to build a metric which leads to perfect precision in determining if a reactor trip is to occur over large timescales (many hours or days), this is largely because many of the complex, interconnected precursors are not known. However

after trip events have occurred, investigations have been completed where the cause of the trip in the preceding minutes or hours are known. Specifically the causes for the trips are known for 11 out of the 15 reactor trips for R1. There are also 15 examples of the data in confirmed *healthy* regions which provide us with a nominal baseline test set. In this section we will look at case examples and then study how we can develop a flagging procedure for potential unexpected trip events. The approach developed involves combining multiple methods outlined in Section 6.3. We run the PCA analysis keeping 1 PCA feature with a rolling 1 week window with on a combination of $\alpha(t)$ analysed features. Naturally, PCA requires a set of features so we cluster them into natural combinations of subsystems. For example we could consider "CC236" clusters which incorporates the following features "R1 CC236 A" and "R1 CC236 B" or alternatively "CC" features will take all condenser features. Clustering the system in "CC" type features we have 14 original features including "CC205", "CC236", "CC237", "CC601", "CC206", "CC213", "CC214" and "CC216" with A and B channels where appropriate.[1] It is also important to note that the metrics and confusion matrices reported are not a perfect indicator of the performance of the models. Much of the data in the *monitor* regions may in fact occur before any fault develops and also anomalous data might exist in *healthy* regions of which we are unaware. Therefore they may be mislabeled data within the dataset, so the metrics reported should be taken largely with a qualitative interpretation not a quantitative one.

Here we will report results for the example *monitor* region (19/7/2019 – 9/10/2019) in which a fault developed which lead to an unresponsive control rod in the hour preceding the trip. The PCA analysis flags 6 regions; Fig. 6.10 shows a screen grab from the analysis dashboard (cf. Appendix B) which indicates the 6 flag regions.[2] For convenience, the following regions are highlighted:

[1] A and B refers to independent subsystems of the feature which are otherwise identical in a given reactor.

[2] The dashboard has a zoom-in functionality to aid the user and reports the number of flag regions.

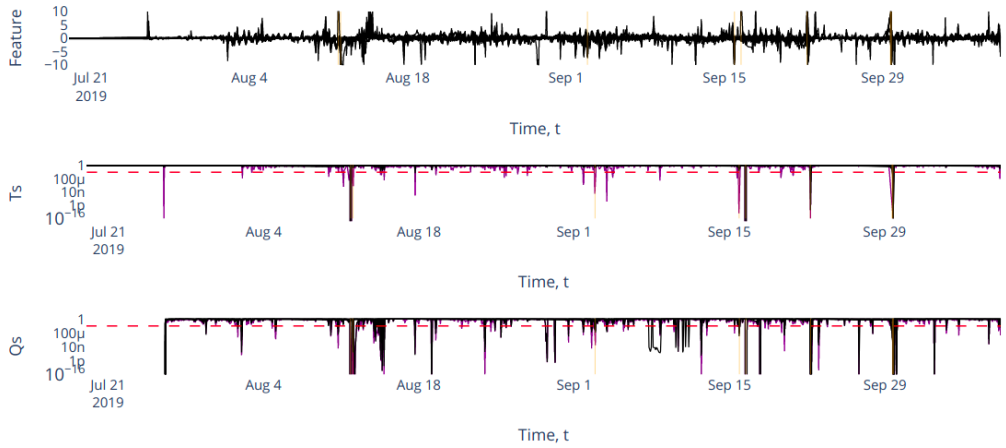


Figure 6.10: Under the hood view of Dashboard PCA statistics. An trip *monitor* region using features of type "CC". Top: Normalised variable (coloured) and overall (black). Centre: p-values of the T^2 statistics for each feature (black) and overall (purple). Bottom: p-values of the Q statistics for each feature (black) and overall (purple). The red lines indicate the α_{crit} confidence limit threshold.

- 12th August 2019,
- 2nd September 2019,
- 16th September 2019,
- 22nd September 2019,
- 29th September 2019,
- 9th October 2019.

A few notes about these regions, 9th October is detecting the initial onset of the reactor trip. The rods were found to be drifting into the core which will alter the majority of the parameters. By looking at the marginalised T^2 and Q statistic we determine that the pressure and temperature in the condenser in the hour preceding are the culprit features. It is expected that the temperature and pressure will be affected if the reaction is being throttled. Fig. 6.11 shows a dip in many of the

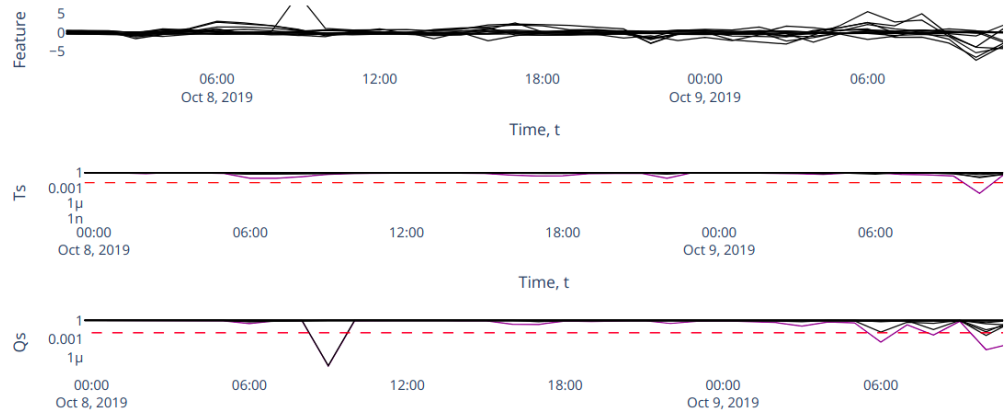


Figure 6.11: Under the hood view of Dashboard PCA statistics. An trip *monitor* region using features of type "CC". Top: Normalised variable (coloured) and overall (black) Centre: p-values of the T^2 statistics for each feature (black) and overall (purple) Bottom: p-values of the Q statistics for each feature (black) and overall (purple). The red lines indicate the α_{crit} confidence limit threshold.

features in the upper plot leading to T^2 and Q dropping below the α_{crit} thresholds at 6am 9th October 2019, the reactor trip time was recorded at 11:36am. An important consideration about the flagged regions is that if we have an impulse in the features we can often see the α - analysis report it twice. This occurs for 22nd September and then 29th September one week later exactly as the impulse leaves the rolling PCA window it notices a change in the distribution which is simply an artefact of the previous anomaly (cf. Fig. 6.12) leaving the window.

In the dashboard we can see which features are responsible for the flagging by taking the marginal contributions of each feature to the test statistics. In Fig. 6.12 the test statistics are dominated by changes in "R1 CC601 B". Taking a look at the raw plots of this feature we see a sharp change in its behaviour. In Fig. 6.13 we show the raw feature data for two of the features and clearly see an impulse in "R1 CC601 B" on the 22nd September, however we see no such large deviation in other features such as "R1 CC237 B" at this date. These are the kind of changes in the behaviour of the data we wish to successfully detect and determine features or subsystems which are

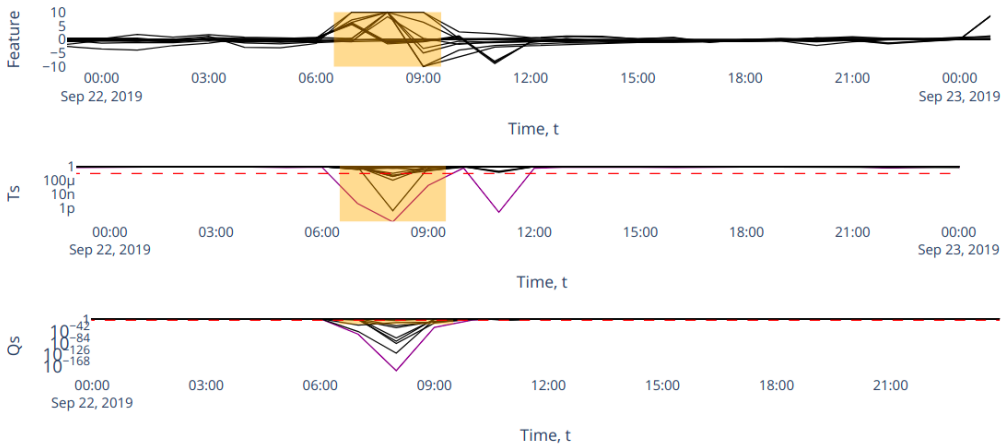


Figure 6.12: Under the hood view of Dashboard PCA statistics. A trip *monitor* region using features of type "CC". Top: Normalised variable (coloured) and overall (black). Centre: T^2 statistics for each feature (black) and overall (purple). Bottom: Q statistics for each feature (black) and overall (purple). The red lines indicate the α_{crit} confidence limit threshold.

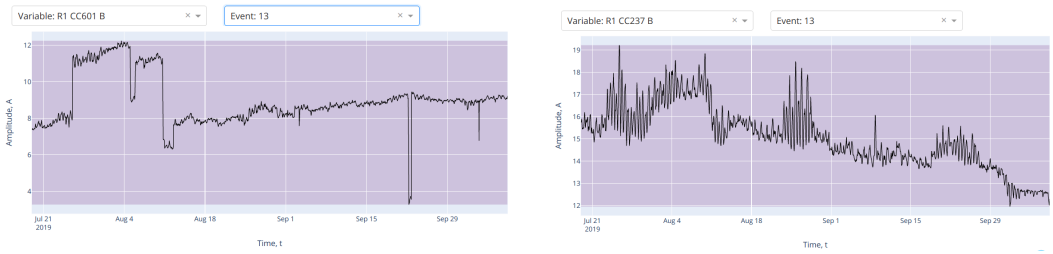


Figure 6.13: Dashboard view of two features.
Left: Dashboard screen grab from variable plots of "R1 CC601 B". Right:Dashboard screen grab from variable plots of "R1 CC237 B".

responsible. The analysis is also unaware of intentional reactor operation changes however an engineer would be aware of how operational changes in the reactor are likely to change measurement readings. We can get a measure of our success by comparing the rate of flagging anomalous regions between the *healthy* control regions and the *monitor* regions as defined in section 6.1. If we profile α_{crit} for the feature types "CC" we produce Fig. 6.14.

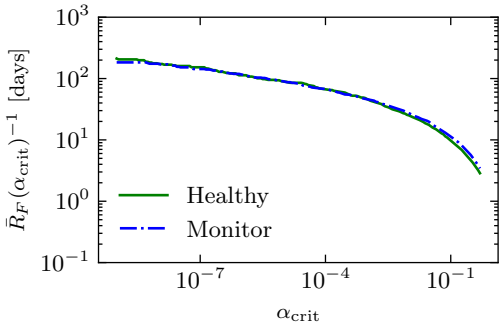


Figure 6.14: α - analysis results for the "CC" family of features. The 1 hour duration flag rates between the *monitor* and *healthy* classes.

In particular for $\alpha_{\text{crit}} = 0.01$ we report an average flagging rate $\bar{R}_F(\alpha_{\text{crit}})$ which describes the time average number of α_{crit} threshold exceeding events over the data:

- *monitor* regions: $\bar{R}_F(\alpha_{\text{crit}})^{-1} = 25.1$ days,
- *healthy* regions: $\bar{R}_F(\alpha_{\text{crit}})^{-1} = 27.3$ days.

The result indicates on average we see more anomalous data points in the regions known to lead up to a trip than those not per day, however the rates are not significant between the two classes. In fact, this margin is off little use for distinguishing possible anomalous behaviour. The sensitivity of the tools can be improved and adjusted by tweaking the α_{crit} parameter if this rate is deemed too frequent or not frequent enough. We introduce $\bar{\beta}_{\text{crit}}$ which exploits the noted difference in flag rate between the classes derived from α_{crit} alone which gives us room to build a classifier. Calculating the rolling mean of flag rates, $\bar{\beta}(t)$ of at least 1 hour duration and using these values we can build a simple classifier that picks an optimal threshold rate of anomalous regions for the optimal α_{crit} using Alg. 2. A simple decision tree is employed with one decision $\bar{\beta}(t) > \bar{\beta}_{\text{crit}}$ then data points within the window $t - T_{\text{window}} : t$ are labeled a monitor region, thus we now report the flagging rate, $\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})$ which describes rolling time average of the $\bar{\beta}_{\text{crit}}$ threshold exceeding events over the data.

In Fig. 6.15 we see a subtle difference in the flag rates between *healthy* and *monitor* data where the *healthy* line is consistently above *monitor* for $\alpha_{\text{crit}} < 0.01$. This

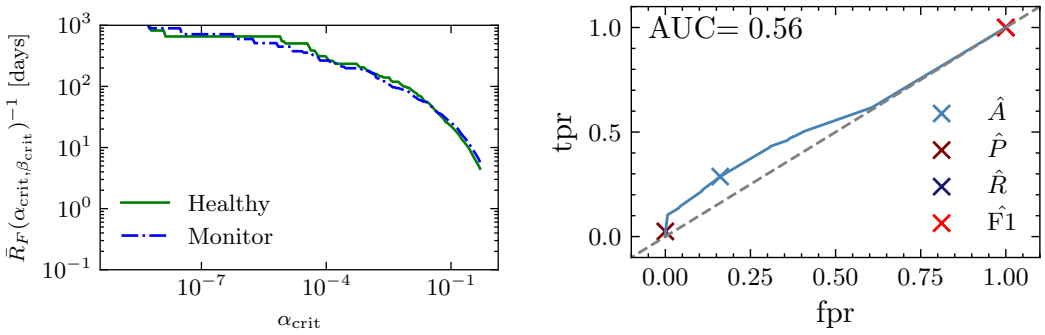


Figure 6.15: α - analysis results for the "CC" family of features.
Left: The 1+ hour duration rolling average flag rates between the *monitor* and *healthy* classes optimised for accuracy. Right: ROC curve, key threshold choices maximising each of accuracy A, precision P, recall R and F1 score are marked with crosses.

indicates anomalous behaviour in the *healthy* regions is less common than the *monitor* regions. This is further demonstrated by a weak performance in the classifier and a low AUC in the given ROC curve. It is important to recall we built a dataset by naively assuming that all the data within a *healthy* region contains no anomalies and a *monitor* region is entirely anomalous. This leads to poor performance metrics due to large proportions of the dataset potentially being mislabelled, so the reported confusion matrices and ROC curves are shown as demonstrations of the classifying power despite this naive approach, thus exploited with the aid of human investigation, but not automation. The dataset could not be labeled more precisely as the exact precursor to the events that led to anomalous or unexpected operational changes in the reactor are not always determined. The results from introducing $\bar{\beta}_{\text{crit}}$ as the classification boundary are presented in Tab. 6.1 for optimisations in A and F1 score.

The rates $\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})^{-1}$ indicate that the models are flagging anomalous data points in the *monitor* regions more frequently than the *healthy* counterparts. This demonstrates an improved classifying power by the introduction of a rolling average flagging rate them even if it is limited. If the choice of optimisation metric is F1

metric	α_{crit}	$\bar{\beta}_{\text{crit}}$	$\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})^{-1}$	
			Monitor	*Healthy*
A	0.18	0.33 day^{-1}	13.5 days	16.3 days
F1	0.015	0.40 day^{-1}	65.0 days	77.0 days

Table 6.1: The "CC" alpha processed dataset results for the rolling average flagging rate analysis.

score we retrieve a much higher $\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})$ than for accuracy. F1 score reduces the number of FN but at the expense of a lower trigger rate over both classes. This could be a suitable trade off where we do not wish the detection system to trigger repeatedly and frequently. We can further dissect the results of the models with the confusion matrices presented in Fig. 6.16. These matrices are generated using the classification of single data points that exceed the threshold $\bar{\beta}_{\text{crit}}$ not regions with multiple concurrent threshold exceeders.

In the accuracy optimised matrices it is clear there is classifying power shown in values of the TN and TP boxes. It may seem like this is not a significant result giving an overall accuracy of 57.2% but it is important to be aware of a few points. This simple threshold classifier is assuming the rate values form independent measurements when in actually this is not the case. The value of each rate depends on any flags over the previous rolling window and the PCA fitted model which again depends on a rolling window. There are large portions of the *monitor* regions of the dataset which are mostly likely to in fact be *healthy* regions in the reactor due to our naive labeling procedure. It is also perfectly possible that there was anomalous behaviour in a *healthy* region in this test set however because these data points were not before an unexpected trip we assume they are all perfectly nominal.

The results here assume that there is no human intervention or oversight, where a human might simply disregard a flag due to a planned change in the operational parameters (e.g. change in grid power demand) of the reactor which may occur more in certain systems therefore there exists a number of FP which can be quickly

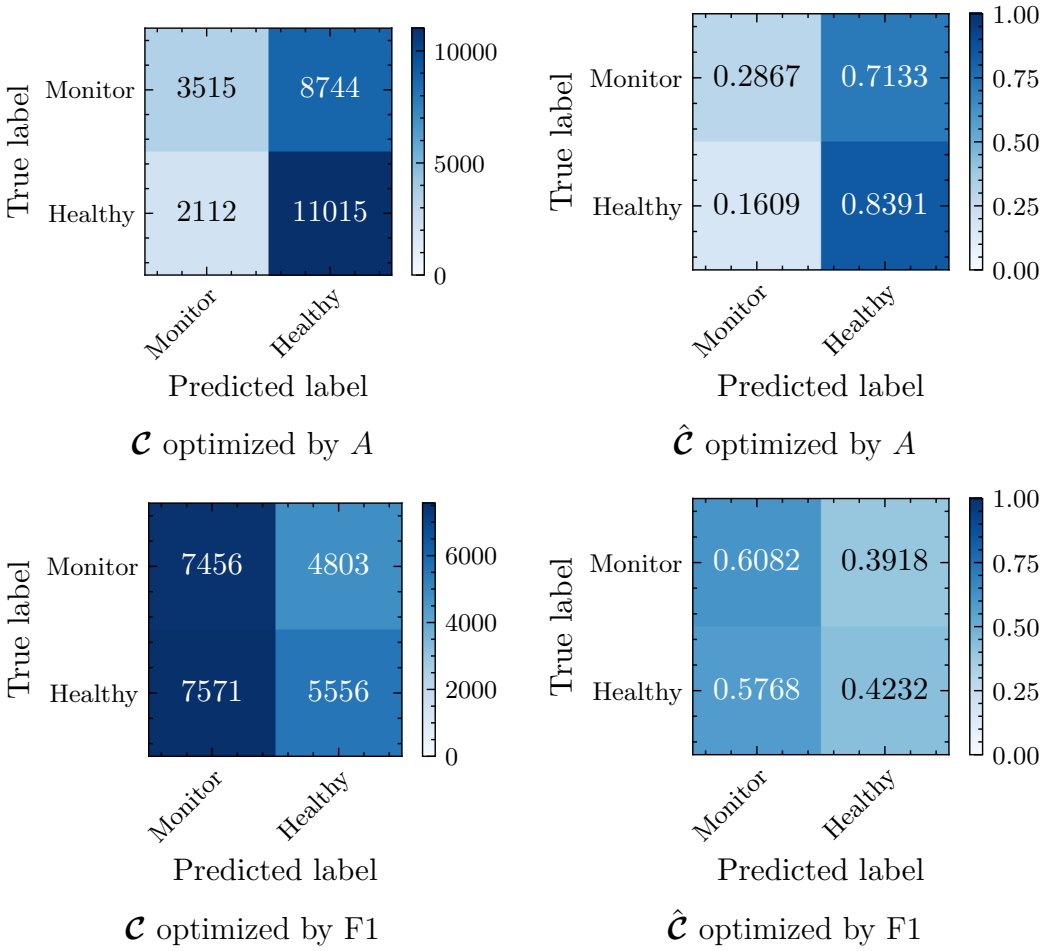


Figure 6.16: The "CC" alpha processed dataset confusion matrices for the rolling average flagging rate analysis. Where $\mathcal{C}$ are the un-normalised and $\hat{\mathcal{C}}$ the normalised confusion matrices respectively.

discarded. Alternatively, an engineer may spot a genuine point of concern and investigate before the PCA models become "thermalised" to anomalous behaviour in the following points. This is why the dashboard has the functionality to look at marginal flag rates responsible from individual features, the user can identify which features have deviated to prompt human investigation or intervention.

These results indicate there is a small level of classifying power in our methods, but it is not suitable for full automation and does require human consideration as opposed to directly taking the classification of the PCA models at face value. Further work into various subsystem of the "CC" class of features could have yielded better results

than the combination we present here. We will now turn our attention to another dataset which is more closely connected to the reactor core itself rather than its supporting systems.

6.5 Channel Gas Outlet Dataset

The methods outlined in this work can also be applied to other datasets including the CGO dataset. This dataset comprises of 324 channels of temperature and gag positions of the control rods over time. These features are strongly correlated with the power output of the reactor so potential good indicators for reactor health. We can use the toolkit we have developed and use the following features; reactor power, gag and thermocouple for each channel in which we aim to find indications of failing thermocouples or unexpected changes in reactor power output. In some cases thermocouples are known to be faulty in otherwise *healthy* operation of the reactor, the large number of thermocouples and known symmetries in the reactor mean this is an acceptable tolerance of operation in the reactor. For example it is known that there exist channel partners in the reactor, which are pairs of highly correlated channels due to the symmetric nature of the reactor which introduce a redundancy in the instrumentation, but for now we omit that consideration (see Fig. 6.17). It is beneficial to find a developing fault such that instrumentation can be replaced in the following reactor refueling cycle to improve efficiency of plant maintenance. In the 324 channels we take an hourly average over each feature in the period 1st January 2018 – 31st December 2019 where the reactor is running at full power. For each channel we know exactly where it was considered "healthy", "degrading" or "faulty" and we discard any data after which the channel thermocouple is considered faulty, therefore our methods can be applied to flag anomalies in the raw data before a region is declared faulty. A "degrading" thermocouple is one which displays a miscalibration or loss of sensitivity but it is still usable whereas a "faulty" thermocouple behaves incorrectly and the readings need to be ignored entirely. Here

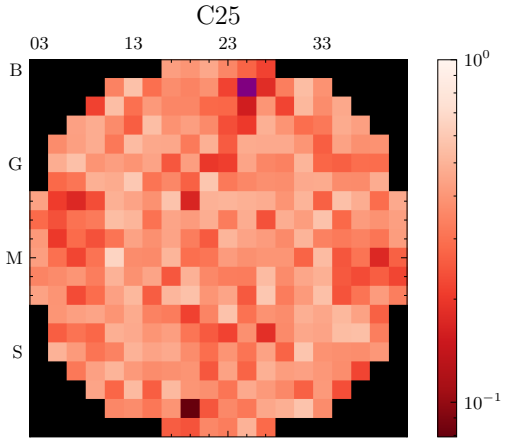


Figure 6.17: Cross section of R1 reactor core indicating the 324 channels, C25 is shown in purple. Heatmap depicts the p-values of the Engle-Granger co-integration test [117] for temperature of channel C25. V19 is found to be highly co-integrated – this is the channel partner of C25 indicated in deep red diametrically opposed to C25.

we use raw feature data and not α - analysis because the CGO dataset does not express any periodicity and the time series is largely stationary. In Fig. 6.18 we have

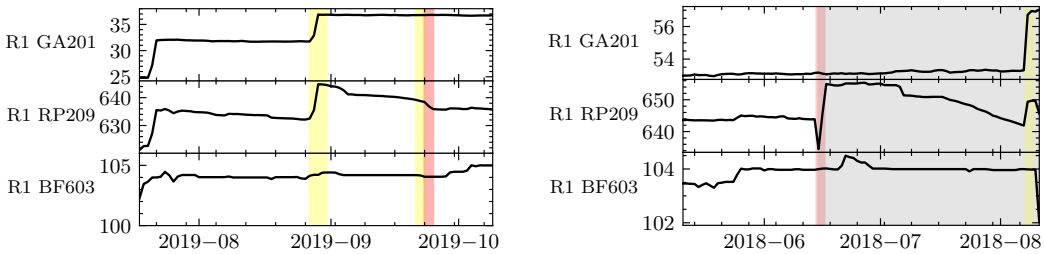


Figure 6.18: Two example datasets of a the CGO correlated features, Gag position "GA201", Thermocouple temperature "RP209" and Power output "BF603". Left: CGO data for channel C25, a channel not regarded as faulty. Right: CGO data for channel M05, a channel regarded as faulty in the grey region onward.

two channels for comparison, the yellow regions indicate any combination of features are violating the α_{crit} threshold and red when the marginalised thermocouple T^2

and Q statistic p-values are dominant. If a marginalised feature is responsible for a flag, then that suggests that there is not a coupling of behaviour in the features in the reactor such as those expected from asking the reactor to reduce the power output by moving the gag positions. A single marginalised feature being responsible suggests a change in instrumentation not reactor behaviour. In C25 there are some regions identified which are explained by adjustments to the gag position or power, the first jump in each is not highlighted because it is contained within our first PCA window (i.e. the thermalisation time) of two weeks duration. In the M05 channel we see in the red highlighted region that the value of the thermocouple shifts but the power and gag do not, this is indicative of a fault. The PCA method has successfully picked out the fault on the same day as it was determined by engineers. In C25 there is also a red region which corresponds to the thermocouple being declared degrading but not faulty in the CGO logs. This method does lead to a high false positive rate, in fact on average in any channel there is one false positive every $\sim$50 days. However correcting for the average expected adjustments to the gag or output over a channel around 3.75 times per year per channel (of which the engineer will be aware) we recover a false positive rate of once every $\sim$154 days per channel, this gives us a greater confidence in the anomaly detection toolkit. There are in fact 4 channels in which a fault develops in 2018 – 2020 and we get at least one flag in the previous month for each channel which is indicative of an anomaly. Now we have justified structural differences in the dataset we apply the PCA rolling window rates method. Here we have 14 *healthy* and 14 *monitor* regions by considering the data from 2010 – 2020 (performing the same grooming of data as the condenser and vacuum data) and perform our rolling average rate analysis on flagged regions which are at least 1 day in length. Similarly to the previous dataset in section 6.6 we can build a simple threshold rates classifier. We still have classifying power despite the apparent rates appearing to be more similar than for the condenser and vacuum dataset (cf. Tab. 6.19). The results from performing optimizations for accuracy and F1 are presented in Tab. 6.2.

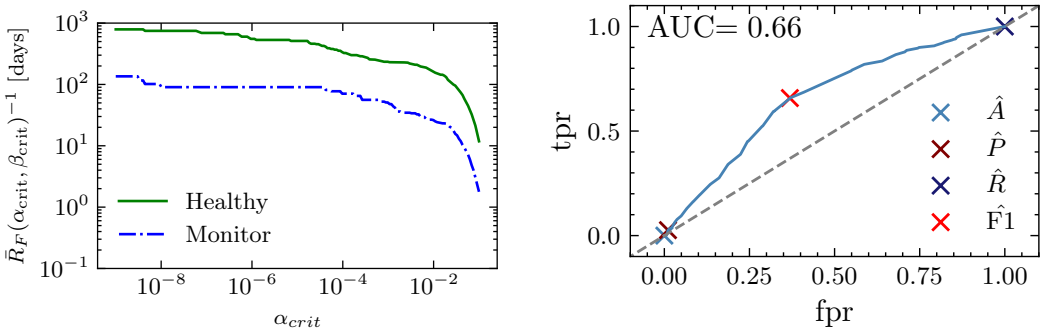


Figure 6.19: Results for the CGO family of features, "RP209", "GA201", "BF603".
Left: The 1+ hour duration rolling average flag rates between the *monitor* and *healthy* classes optimised for accuracy. Right: ROC curve, key threshold choices maximising each of accuracy A, precision P, recall R and F1 score are marked with crosses.

metric	α_{crit}	$\bar{\beta}_{\text{crit}}$	$\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})^{-1}$	
			Monitor	*Healthy*
A	0.41	0.071 day^{-1}	∞ days	∞ days
F1	0.45	0.38 day^{-1}	10.3 days	66.0 days

Table 6.2: CGO dataset results for the rolling average flagging rate analysis.

These results demonstrate clear power in distinguishing between the *monitor* and *healthy* classes with a significant differences between the $\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})^{-1}$ values for the F1 score optimisation. The optimisation for accuracy in this case clearly demonstrates why it is a poor metric for imbalanced classes. The best accuracy score for this data is found when the model chooses an unachievable $\bar{\beta}_{\text{crit}}$ value of 0.071 day^{-1}. This value is larger than the rolling window size therefore $\bar{\beta}_{\text{crit}}$ never exceeded and we never predict a *monitor* region. These results are presented in a set of confusion matrices in Fig. 6.20.

We get considerably more balanced confusion matrices here than the condenser and vacuum datasets looked at previously. This dataset requires more careful consideration than the "CC" features as there are frequent adjustments to the gag positions

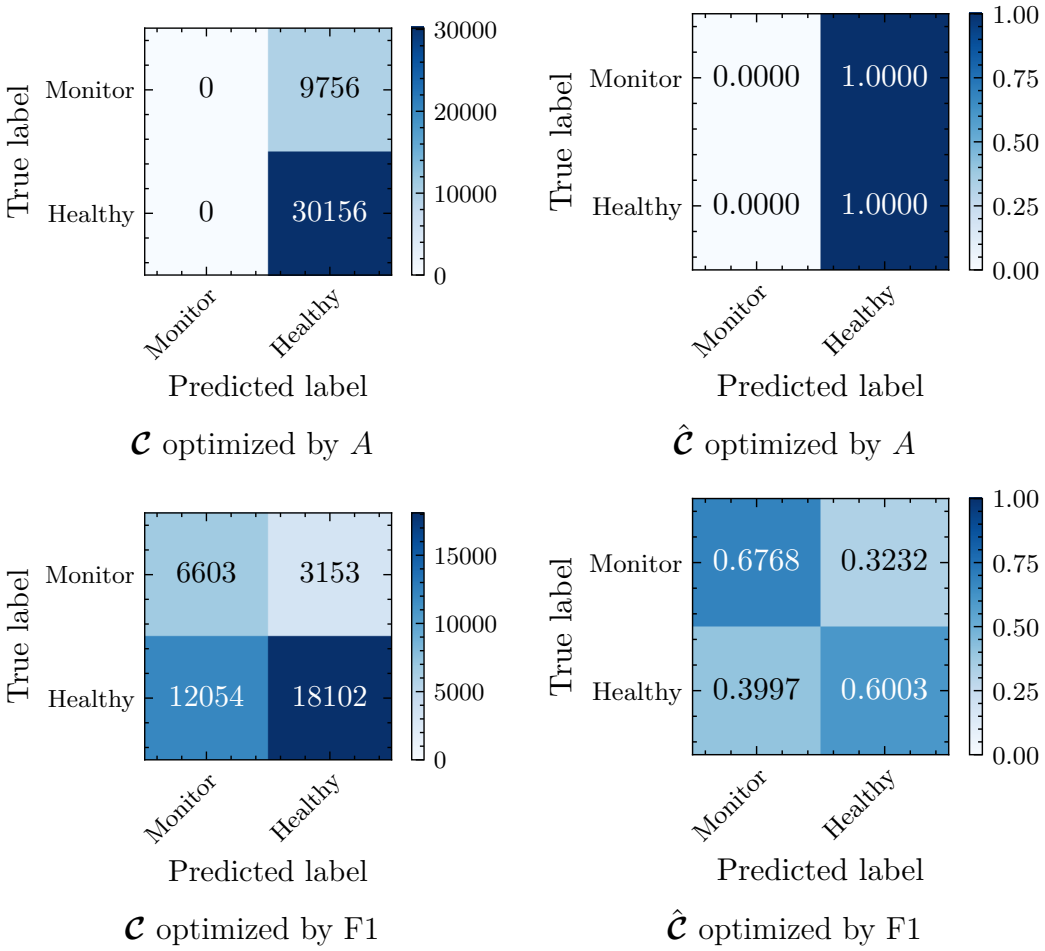


Figure 6.20: The CGO dataset confusion matrices,for the rolling average flagging rate analysis. Where $\mathcal{C}$ are the un-normalised and $\hat{\mathcal{C}}$ the normalised confusion matrices respectively.

and power outputs due to operational changes due to power grid demands or reactor fuel stages. The many operational reactor changes are easily identifiable by an engineer so in many cases these events can be discarded immediately. Eliminating these changes from the dataset however provides more promising results for the use of the PCA toolkit in the CGO dataset than the “CC” family of features.

6.6 Conclusions

The tools developed and discussed in this work are intended to be entirely informative for an engineer and makes no claim at predicting the possibility of an unexpected reactor trip in a certain period of time. What the tools do is provide a way to examine the distributions together through PCA and separately through BOCPD to provide an engineer with greater insight into potential subtle changes in the operational parameters of the reactor systems which could be indicative of faults developing or unexpected changes. The BOCPD is the most powerful, for efficient and easy interpretability of the predictive of changes in the distributions across all the provided features, it also had the additional bonus of being easily adapted to online detection. However much like the PCA a certain amount of preprocessing is required to make the data stationary which is performed by the α-analysis and first order differencing. These methods unfortunately sacrifice a window of data for model thermalisation. The PCA models provide another perspective into how well we can understand the live data against fitted models and conversely how anomalous the data is. The average PCA is limited in the sense that anomalous points are absorbed into the previous window fit, this can make PCA less sensitive to other changes in the window as the variance in window has been increased.

The success of the models is demonstrated by examining the rate of flagging regions, which in themselves can be used to build a classifier. While the results reported may not be perfectly labeled as we have made assumptions about individual set of data points in each class, the results do demonstrate clear classifying power in our methods. These methods if implemented could be further validated by tackling the naive data labeling procedure. This would allow the methods to be refined with a more appropriate dataset while also reporting correct metrics of performance. We find the optimal threshold p-values, (α_{crit} and $\bar{\beta}_{\text{crit}}$) for accuracy or F1 metrics and if the rolling rate exceeds this value there is an indication that the systems are developing anomalous behaviour. These thresholds have been fitted for a selection

of features and implemented into a dashboard with a user guide which can be found in Appendix B.

Chapter 7

Social Mixing Matrices at Cox's Bazar

7.1 Introduction

Contact matrices are an important ingredient in age-structured epidemic models to inform the simulated spread of the disease between sub-groups of the population. These matrices are generally derived using resource-intensive diary-based surveys and few exist in the Global South or tailored to vulnerable populations. In particular, no contact matrices exist for refugee settlements – locations under-served by epidemic models in general. In this section we present a novel, mixed-method approach, for deriving contact matrices in populations which combines a lightweight, rapidly deployable, survey with an agent-based model of the population informed by census and behavioural data. We use this method to derive the first set of contact matrices for the Cox's Bazar refugee settlement in Bangladesh. To validate our approach, we apply it to the UK population and compare our derived matrices with well-known contact matrices collected using traditional methods. Our findings demonstrate that our mixed-method approach addresses challenges faced by traditional and agent-based approaches to deriving contact matrices. It also shows potential for implementation in resource-constrained environments.

Epidemics such as COVID-19 have led to devastating consequences for afflicted individuals and their societies. Understanding how such infectious diseases spread, anticipating future trajectories for transmission and gathering evidence to inform decision-making efforts to prevent, mitigate and respond to epidemics is therefore of vital importance. Mathematical and computational models to simulate disease spread are regularly used to support these efforts. Contact matrices are key to understanding social mixing patterns in populations and a vital input to epidemiological models [43, 44]. Despite renewed efforts to develop such models, additional work must be done to ensure they are available to all [118].

In this work we present a new method for determining contact patterns based on combining the information gained from increasingly sophisticated models of disease spread, with that from lightweight surveys which can be rapidly rolled out to populations of interest. We attempt to provide information on contact patterns without requiring the traditional, costly methods of contact data collection. Specifically, we will focus on the use case of the Cox's Bazar refugee settlement in Bangladesh. Epidemics in refugee and internally displaced person (IDP) settlements are commonplace and tend to spread rapidly [119] and only very few models have been designed to simulate outbreaks in these unique environments and to inform public health decision-making [118]. Given the application domain, we believe this is not just an important area in which to contribute knowledge about disease spread patterns, but also a challenging test case which demonstrates the strengths of our methodology.

Throughout this work we will use the JUNE-COX model [42], an agent-based model built on the JUNE framework [23]. An agent-based model is one in which individuals are modelled stochastically with unique properties. In the JUNE model we construct a virtual population at the level of individual residents within a digital twin of the Cox's Bazar settlement, where the individuals have a sex and age.[1] Interactions are simulated between the agents – the virtual residents – in a number of "venues"

[1]Individuals also have an infection history, which we will not consider here except to state that the JUNE-COX model is fitted with epidemic responses including policies like isolation and hospitalisation. These results run with simulations in which all disease models are turned off.

or "locations" that include: shelters; food distribution centres; market places; and learning centres. We use the information from the lightweight survey to guide these interaction patterns based on the demographics of the agents attending the venues over time.

The contact matrices encode information on the number and duration of contacts between people of one age group and another and are usually specific to certain venues or locations in which people interact. There are various types of matrices which can be used both separately and in combination, including (i) one-directional, contact matrices NCM [45] which count the (normalised) number of contacts a person in category i has with a person in category j, (ii) bi-directional reciprocal matrices $\text{NCM}_{\mathbf{R}}$ [45] which also add the number of contacts people in category j have with persons in i and (iii) venue contact matrices $\text{NCM}_{\mathbf{V}}$ [46, 47] which assume that every person at venue L has contact with everybody else present. Here we will discuss a approaches for estimating all three types of matrices. Traditionally, contact matrices are derived using large scale surveys in which participants record the number of contacts they have in different locations and the ages of the people they came into contact with. Additional metadata is sometimes collected, such as the intensity of the contact (e.g. physical or non-physical) and the duration of each individual contact. Surveys of these types have predominantly been run in the Global North, with comparatively few serving countries in which many particularly vulnerable communities reside [120]. One such work has published contact matrices for an IDP settlement [121] and no such work exists on contact matrices in refugee settlements. While such traditional methods of collecting contact data may be considered the gold standard, they are extremely resource consuming to collect and therefore cannot be run easily during an ongoing outbreak. As an alternative to these expensive direct means of contact data collection, several other methods have sought a more indirect approach. Using the information from existing contact surveys conducted in 8 European countries [45] and knowledge of the underlying demographic structures in these populations, Prem *et al.* [122] used a Bayesian hierarchical model to project

these matrices onto 144 countries given similar demographic data and underlying similarities between each of these countries and the original 8 selected in the direct data collection. This has recently been expanded to 177 countries [123].

Similarly, census/demographic data have also been used to construct synthetic populations which are then used to estimate contact matrices. Fumanelli *et al.* [46] use such data from 26 European countries to construct representative synthetic household, school, workplace and 'general community' environments and then assume that each individual in each setting has a single contact with every other member. This has been extended to 35 countries, while also incorporating finer-grained data to develop more representative virtual populations [124]. The same approach is used by Xia *et al.* [125] for the setting of Hong Kong. While such approaches are beneficial as they do not require the expensive collection of long-term contact survey data, they are limited by the assumption that different venues contain static populations and that within venue mixing is homogeneous.

By combining demographic data with data sources such as time use surveys [126,127] or transportation surveys [47], stochastic approaches – e.g. agent-based models – have been developed to capture a broader variety of mixing patterns in populations. These approaches expand on those described above by exploring many permutations of possible venue mixing. Despite this, these methods still present similar limitations to those described above. Namely, in the absence of any prior information on interaction patterns, it is largely assumed that each agent contacts every other agent in those venues. As a partial remedy to this challenge, disease data is commonly used to fit arbitrary hyperparameter multipliers to these matrices. While this is generally a necessity to be able to forecast disease spread even when using directly collected contact data [128], due to differences between disease transmission routes this may not resolve the errors at the matrix element level. The output of this process does not provide an understanding of the base level of contacts, but rather a set of contact matrices for each disease. This limits the usefulness and generalisability of such matrices in comparison to corresponding matrices from directly collected data.

In this work, we seek to accomplish two things: i) Develop a methodology which addresses the challenges above by taking a mixed-method approach to deriving contact matrices. It combines techniques of extracting contact matrices from sophisticated agent-based models, with information derived from a lightweight survey designed to inform and validate the model-derived matrices, while being significantly less expensive to run than the traditional large-scale contact surveys. ii) Use this new approach to present some of the first contact matrices for a refugee settlement. Because of their use in different types of models the matrices need different normalization, either to the full population, as in the case of location-unspecific simple compartment models of the SEIR (susceptible, exposed, infected and recovered) type, or to the part of the population actually visiting a venue. We will therefore present results for all three types of contact matrices, for a variety of locations, either normalised to the overall population "P" type contact matrices (PNCM, $\text{PNCM}_{\mathbf{R}}$ and $\text{PNCM}_{\mathbf{V}}$) or to the actual users of a location "U" type matrices (UNCM, $\text{UNCM}_{\mathbf{R}}$ and $\text{UNCM}_{\mathbf{V}}$).

This work was motivated to contribute to the global call to action laid out in prior work, which aims to develop new methods and mechanisms of data collection for modelling disease spread in refugee and IDP settlements [118].

7.2 Methods

The goal of our method is to construct location-dependent social contact matrices with a high level of granularity without resorting to detailed contact surveys. We achieve this by fitting the (virtual) contact matrices of an individual-based model constructed from higher-resolution demographic data of the population to the real-world results from lightweight surveys with a much lower resolution. The resolution and accuracy implicit to the model allows us not only to infer the highly-granular contact matrices, but also allows us to give a first estimate of the associated uncertainties. We describe the model as high-resolution here because it is well informed compared to the contact survey. Collection and distribution of aid around the camp and its

demand reliably inform the census data in the camp. However, as the lightweight survey is conducted such a small scale (∼300 people) it has potentially large bias and great uncertainty which take into account. In the following we further detail this procedure and apply it to the construction of social contact matrices for the residents of Cox's Bazar refugee settlement.

7.2.1 The Survey

The level of detail accessed by surveys in refugee camp settings is often heavily constrained by resource considerations (timing, number of enumerators, need for rapid results etc.) and the highly aggregate contact survey run in the Cox's Bazar refugee settlement between October-November 2020 is no exception. During this period, the settlement was continuing to experience cases of COVID-19 [129]. However, reported case numbers were low and the settlement activity had largely returned to pre-pandemic levels, with the exception that learning centres (schools) remained closed and masks were still being worn [130, 131]. The following demonstrates the ability to rapidly run a survey during a public health emergency, in a resource-light way, while producing representative results of the contact patterns which can be used in future studies and modelling works. Although a more intensive survey – such as a diary-based longitudinal study – would provide more precise and accurate data, the ability to perform such a survey may be limited by the number of researchers available or more practical concerns such a limiting social contacts between members of the community and enumerators during a public health crisis.

The survey underpinning the study was conducted by experienced enumerators from the UNHCR Community Based Protection (CBP) team, following standard UNHCR practices [132, 133]. Its objective was to collect information on the number of contacts people of different demographics have with others in different venues they attend during a typical day. The survey considered only three categories of residents, defined by their age: children (< 18 years), adults (≥ 18 and < 60 years)

and seniors ($\geq$ 60 years) and we constrained the set of surveyed locations to those contained in the digital twin, JUNE-COX. Data was collected from 22 camps in the Kutapalong-Balukhali Expansion Site (part of the Cox's Bazar refugee settlement). In each camp the respondents were two male and two female residents in each of the three age brackets. In addition, two persons with disabilities were surveyed in each camp, resulting in a total of $22 \times 14 = 308$ respondents. Details of the survey can be found in Appendix C.2 and the accompanying metadata to the anonymous results [134]. The respondents were asked if they attend various venues and, if so, to estimate the number of adults and children they come into contact with there. To avoid skewing results through uncharacteristically long or short times at a venue, the respondents were asked how much time they generally spend at those venues at any given visit such that the total contacts can be re-scaled to contacts per hour. Since the JUNE modelling framework and much of the demographic data underpinning JUNE-COX do not distinguish adults (18-59) and seniors (>59) we choose to combine the data in these two age bins into one "adult" category, thereby arriving at highly aggregate 2×2 total contact contact matrices t_{ij}[1] for the various locations L.[2] We use the survey to calculate $\text{UNCM}_{\mathbf{R}}$ type matrices for different locations. Here we present the methodology to calculate the different versions of the contact matrices:

1. One-directional contact matrices [45], NCM, (UNCM and PNCM): Following the notation in [135] the PNCM are denoted as M with elements m_{ij} defined by $m_{ij} = t_{ij}/n_j$ with t_{ij} the aggregate total number of contacts of n_j survey respondents in category j reported with people in category i.

   There is a subtle difference to the UNCM with elements μ_{ij}, where the aggregate number of contacts t_{ij} is normalised to the number of actual users in the venue, η_j, $\mu_{ij} = tij/\eta_j$.

   To make contact between the PNCM and UNCM, one therefore merely has

[1] We interpret contact matrix Δ_{ij} such that person i contacts person j and graphically as subgroup on x-axis contacts subgroup on y-axis.

[2] To improve the readability of the manuscript we refrain, where possible, from explicitly indexing contact matrices etc. with a location index.

to re-normalise to the overall number of respondents in category j, $m_{ij} = t_{ij}/n_j = \mu_{ij}\eta_j/n_j = \mu_{ij}a_j$, where a_j denotes the attendance rate to the venue in category j. This re-normalisation can be performed for any conversion from population normalised "P" to user "U" normalised matrices.

2. Bi-directional, reciprocal contact matrices [45], $\mathrm{NCM}_{\mathbf{R}}$, ($\mathrm{UNCM}_{\mathbf{R}}$ and $\mathrm{PNCM}_{\mathbf{R}}$): Following, again [135], the $\mathrm{PNCM}_{\mathbf{R}}$ are denoted by C and their elements are defined as

$$c_{ij} = \frac{1}{2}\left(m_{ij} + m_{ji}\frac{w_i}{w_j}\right) = \frac{1}{2w_j}\left(t_{ij}\frac{w_j}{n_j} + t_{ji}\frac{w_i}{n_i}\right), \tag{7.2.1}$$

where the $w_{i,j}$ are the overall population sizes in categories i and j. This motivates the notion of these matrices being normalised to the overall population. While using these matrices in compartment models, their application in individual-based models may lead to unwanted results. As an example consider the case of contacts between adults and children in school settings and assuming that this is meant to primarily capture the contact of teachers and pupils. Normalising the number of contacts to the overall adult population size would obviously lead to a massively reduced average number of contacts compared to a more correct normalization to the number of teachers in the respective age bins. We therefore define the user-normalised contact matrices $\mathrm{UNCM}_{\mathbf{R}}$ Γ with entries

$$\gamma_{ij} = \frac{1}{2\omega_j}\left(t_{ij}\frac{\omega_j}{\eta_j} + t_{ji}\frac{\omega_i}{\eta_i}\right), \tag{7.2.2}$$

where $\omega_{i,j}$ denote the actual users attending the venue, i.e. $\omega_i = w_i a_i$. In fact, since we resolve the random movement of individuals to distinct locations in JUNE, we use the Γ instead of the C that are more relevant for compartment models. However, we also present results for the population-normalised $\mathrm{PNCM}_{\mathbf{R}}$, which can be obtained by simple re-scaling by attendance factors a_i and a_j from the Γ.

3. Isotropic venue contact matrices, $\mathrm{NCM}_{\mathbf{V}}$, ($\mathrm{UNCM}_{\mathbf{V}}$ and $\mathrm{PNCM}_{\mathbf{V}}$): due to the lack of attendance data we cannot directly derive such matrices b_{ij} and β_{ij} from the survey. However they can be determined virtually.

There is one further subtlety, the contact survey was conducted such that there is an equal number of men and women respondents. However the Cox's Bazar is 48% and 52% men and women by population respectively, we therefore re-weight the contact matrices by sex. Finally, we comment on our treatment of the uncertainties in the survey results. Given the small survey sample size, we right-censor the data at the level of the 90th percentile and perform a bootstrap analysis [136] to determine the median number of contacts between subgroups, μ_{ij}. We assume the uncertainty of this value, $\Delta\mu_{ij}$, to be well estimated by the standard error of the bootstrap distribution of the median. A note should be made that the contact survey is done retrospectively so we expect an element of recall bias within the respondents answers. We choose to right censor the distribution of numerical answers to remove more extreme responses and further we take the median statistic and apply a bootstrap method to estimate median contact patterns and its errors. These errors can be used within our model to apply further stochasticity to agents contact patterns. From $\Delta\mu_{ij}$ it is straightforward to derive the uncertainty, $\Delta\gamma_{ij}$, of the reciprocated matrices, we assume the error in the contacts are dominated by the error from reported number contacts per hour at a venue. We take $\omega_{i,j}$ in Eq. (7.2.2) as an exact quantity from the survey.

7.2.2 The Model

For the construction of the digital twin and simulator we use an existing individual-based model, June-Cox [42], specifying the original June modelling framework [23] to the demographics of the Cox's Bazar refugee settlement. (Note that the original application of the June framework was to model the spread of COVID-19 in the UK and we will refer to this UK specific specification as June-Uk.) Both June-Uk

matrix	symbols	matrix	symbols
CM	t, t_{ij}		
UNCM	μ, μ_{ij}	PNCM	M, m_{ij}
$\text{UNCM}_{\mathbf{R}}$	Γ, γ_{ij}	$\text{PNCM}_{\mathbf{R}}$	C, c_{ij}
UNCM_{V}	β, β_{ij}	$\text{PNCM}_{\mathbf{V}}$	B, b_{ij}
Population venue	η_{ij}	Population world	ω_{ij}
Population survey venue	n_{ij}	Population survey world	w_{ij}

Table 7.1: The CM are time normalised, the various UNCM are further normalised by population at the venues, while the PNCM are instead normalised by the *total* population.

and JUNE-COX use census data to create a virtual population at the individual level, with JUNE-COX specifically focusing here on the Kutapalong-Balukhali Expansion Site of Cox's Bazar. The census data of its population is organised according to a geographical hierarchy; the $\sim$600,000 residents are distributed over the 21 camps ("regions") which make up the Kutapalong-Batukhali Expansion Site (in reality there are 22, however, we combine Camp-20 and the Camp-20 extension together given data availability constrains), these contain between 2-7 UNHCR Admin level-2 blocks ("super areas") comprising $\sim$5000 people, which in turn are composed of sub-blocks ("areas") with 90 households on average. The geographical distribution of individuals and their households is explicitly incorporated in the model through the geo-locations of the area centres.

The sex and ages of the individuals in Cox's Bazar is known to the level of super areas precisely, however we do not know the exact household location of individuals to households. Therefore households are constructed stochastically by clustering

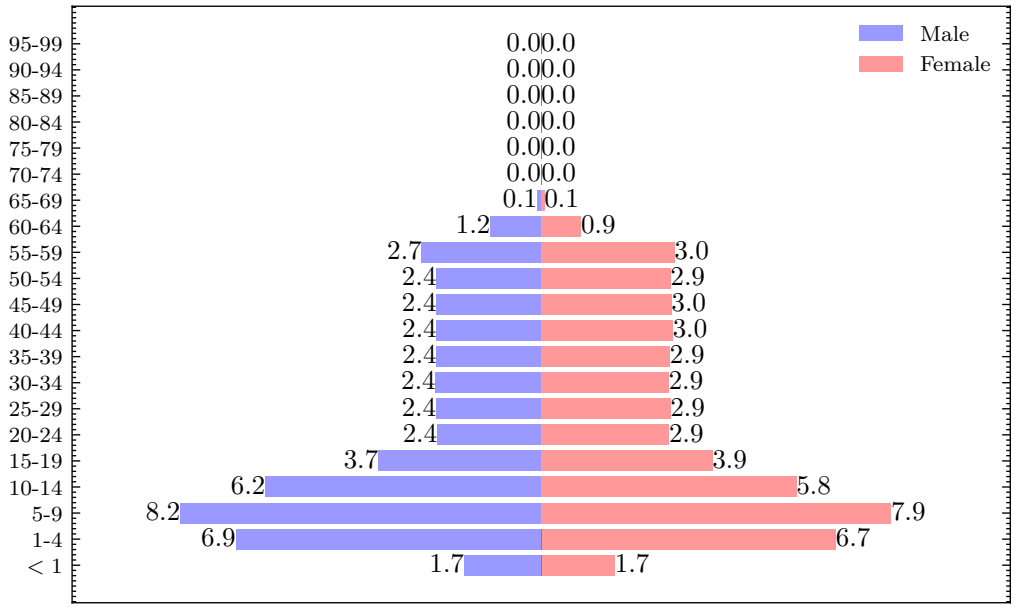


Figure 7.1: The population pyramid of Cox's Bazar. Male population is shown on the left in blue and female on the right in red. The percentage of the population of each category is quoted.

individuals into households according to their age, sex and the following reported properties of the camp in order to create realistic demographic household structures:

- Macroscopic properties:
  1. The distribution of household sizes in the camp (known at the region level) (cf. Fig. 7.3),
  2. The distribution of households in shared shelters (cf. Fig. 7.3),
  3. Population demographics (cf. Fig. 7.1),
  4. The proportion of one, two and multi-generational households (cf. Fig. 7.2).
- Microscopic properties:
  1. The probability of single parent household (cf. Fig. 7.2),
  2. The mean spousal age gap (cf. Fig. 7.4),
  3. The mean age of mother at birth of first child (cf. Fig. 7.4),

4. The distribution of number of children per household (cf. Fig. 7.4).

These properties are all known at the super-area level unless specified otherwise. The resulting household demographic structures can be seen in Fig. 7.2 and shelter sizes in Fig. 7.3. The age brackets for each demographic are inferred from survey and data from the settlement. Children [0 – 18], 18 is the age at which marriage is legal for women (21 for men), Adults [18 – 49] (49 being chosen to provide a realistic age gap for potential grandparents, twice the average mother-child age gap, 22.43 years plus the average spousal age gap, 4.73 years). [49 – 100] for old adults, the remaining ages in the camp. JUNE-COX has an over clustering of children with single parent housing, this due to any remaining children being randomly clustered into households with adults after the children with couples houses are constructed. The microscopic properties of the clustered households are summarised between Fig. 7.2 and Fig. 7.4.

The number of households are predetermined at the area level throughout the camp, however the household demographic statistics are reported at the region level and the household size at the camp wide level. Therefore we have to assume that the camp to region to area statistics can be applied from large geographical areas to smaller ones where the information is unavailable without loss of generality. The agent household clustering is adapted from JUNE-COX [42] as outlined in Appendix S1 and the new methodology is outlined below:

- We first create empty houses with sizes fitting the distribution shown in Fig. 7.3.
- We partition the houses in two groups containing subgroups of demographic properties:
    1. Households with children:
        - Single parent households,
        - Multi-generational households,
        - Two parent households.

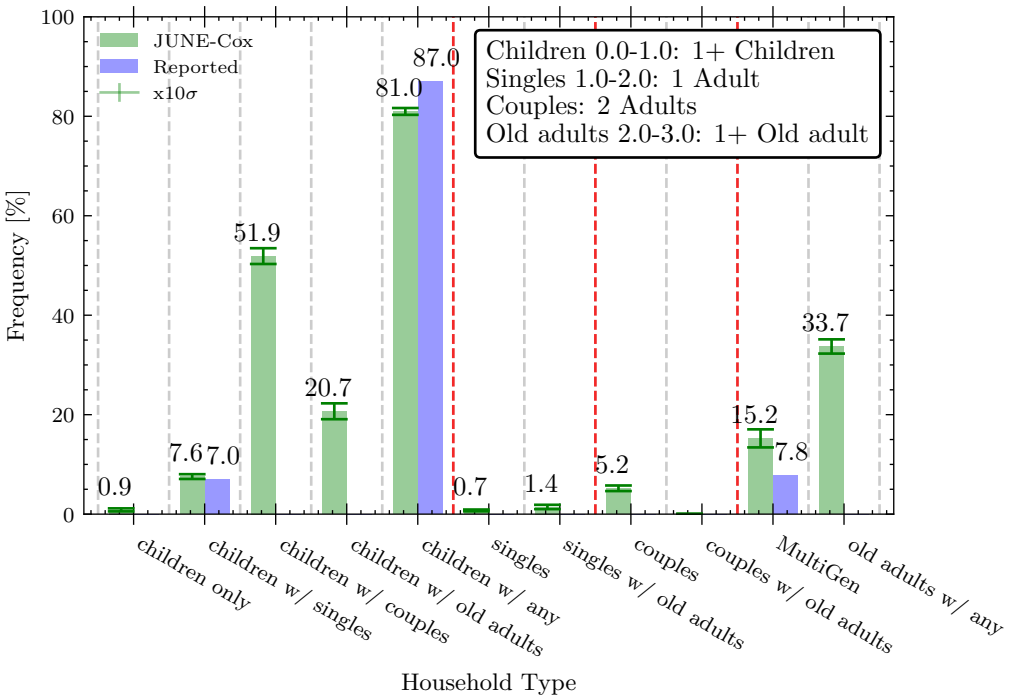


Figure 7.2: Figure of proportion of household types. Note that not all of these groups are mutually exclusive. Green represents the reconstruction in JUNE and Blue the reported data (if available). Those groups where data was unavailable are reported in the figure for completeness. JUNE-COX data reported with scaled error bars of the standard deviation over 20 independent household clustering.

2. Households without children,

- For all households, allocate one young adult (18-49) to each household, if available.

- For single parent households, allocate one young adult (18-49) with an age gap drawn from the adult age gap distribution (see Fig. 7.4) of opposite sex to current resident, if available.

- For all households, allocate all children into households randomly while enforcing a 1 year age gap between them and adult child age gap drawn from distribution (see Fig. 7.4), if available.

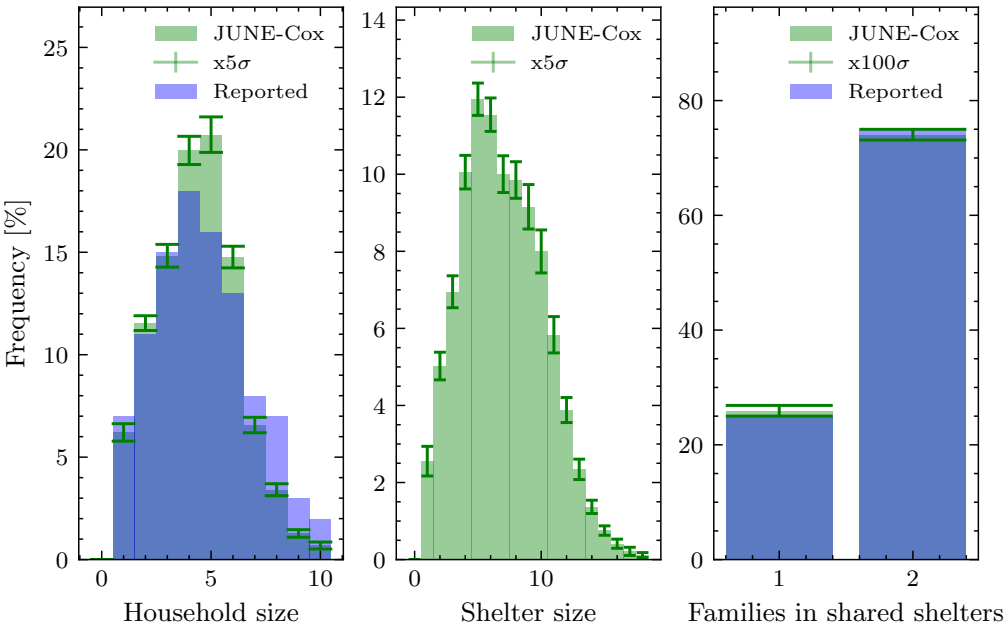


Figure 7.3: Figure of key shelter properties. Left: distribution of household family sizes. Middle: distribution of shelter sizes. Right: Proportion of one and two household shelters. Green represents the reconstruction in JUNE and Blue the reported data (if available). JUNE-COX data reported with scaled error bars of the standard deviation over 20 independent household clustering.

- Allocate older adults (50+) to multi-generational households with an adult age gap or adult child age gap drawn from distributions (see Fig. 7.4) for opposite sex old adult current resident (grandparents age gap) and young adult current resident (parent – grandparent age gap), if available.
- Any remaining adults (18+) are allocated randomly to any remaining households with space under the following priority,
    1. Multi-generational households,
    2. Households without children,
    3. Households with children.

This methodology ensures that JUNE-COX households reflect the household demo-

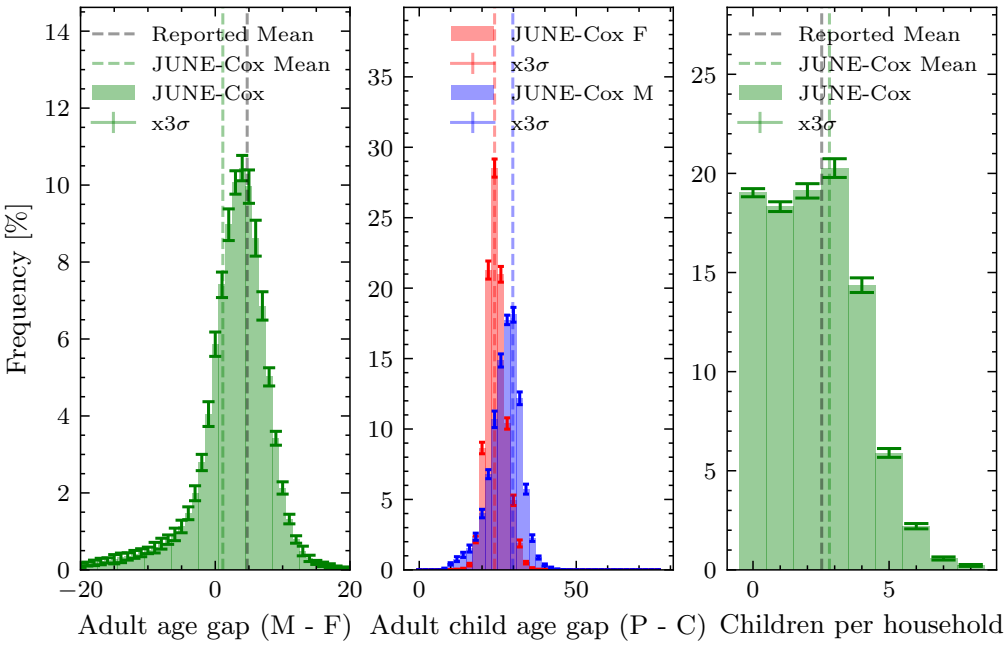


Figure 7.4: Figure of microscopic household properties. Left: distribution of household male female age gap in houses containing only two adults in the range [24-49]. Middle: distribution adult child age gaps in households containing one adult [24-49] and the eldest child [0-18]. Right: distribution of number of children by household. Green represents the reconstruction in JUNE-COX and black dashed lines the reported mean data (if available). JUNE-COX data reported with scaled error bars of the standard deviation over 20 independent household clustering.

graphics with a key emphasis on ensuring multi-generational households are represented proportionately.

After the individuals are created and clustered into households, JUNE-COX constructs different venues in the settlement given their latitude and longitude coordinates: food distribution centres; non-food distribution centres (including LPG distribution centres); e-voucher outlets; community centres; safe spaces for women and girls; religious centres; learning centres; hand pumps and latrines. To simulate the movement of individuals in the settlement we decompose each calendar day into discrete time-steps in units of hours. JUNE uses calendar days to distinguish weekday and

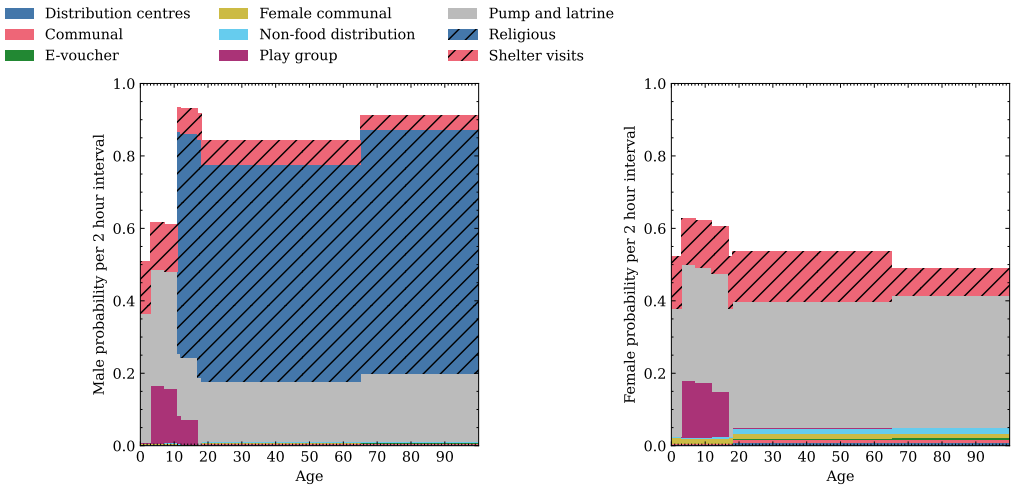


Figure 7.5: The mean probability to attend certain venues in any weekday 2 hour timestep interval by age by sex. Left: Men. Right: Women.

weekend activity profiles where certain venues will be closed.[1] Many individuals have fixed, static, activities, such as the 4 hours at the learning centres for enrolled children and the adults specified as teachers. There is also a fixed 14 hours night-time period, during which everyone returns to their shelter. However, the remaining time is free and people are distributed dynamically. Each person not otherwise occupied (e.g. working, or at a medical facility) is assigned a set of probabilities for undertaking other activities in their free time in the model. These probabilities are part of our social interaction model and depend on the age and sex of the person (Fig. 7.5). They are based on previously collected data capturing daily attendance rates and coarse estimates in proportions of adult/child and male/female attendance (see previous work for details on these calculations and associated data sources [42] and have been further augmented by a series of interviews with CBP officials as detailed in Appendix C.3 and probabilities tuned such that populations reflect the interview see Fig. 7.6).

Given N possible activities with associated probabilities per hour given by $\lambda_1, ..., \lambda_N$, for an agent with characteristic properties p, the overall probability, P, of an in-

[1]This can be further customised in JUNE to specific daily closures or behavioural patterns, however for this proof of concept we only specify weekday and weekend variations.

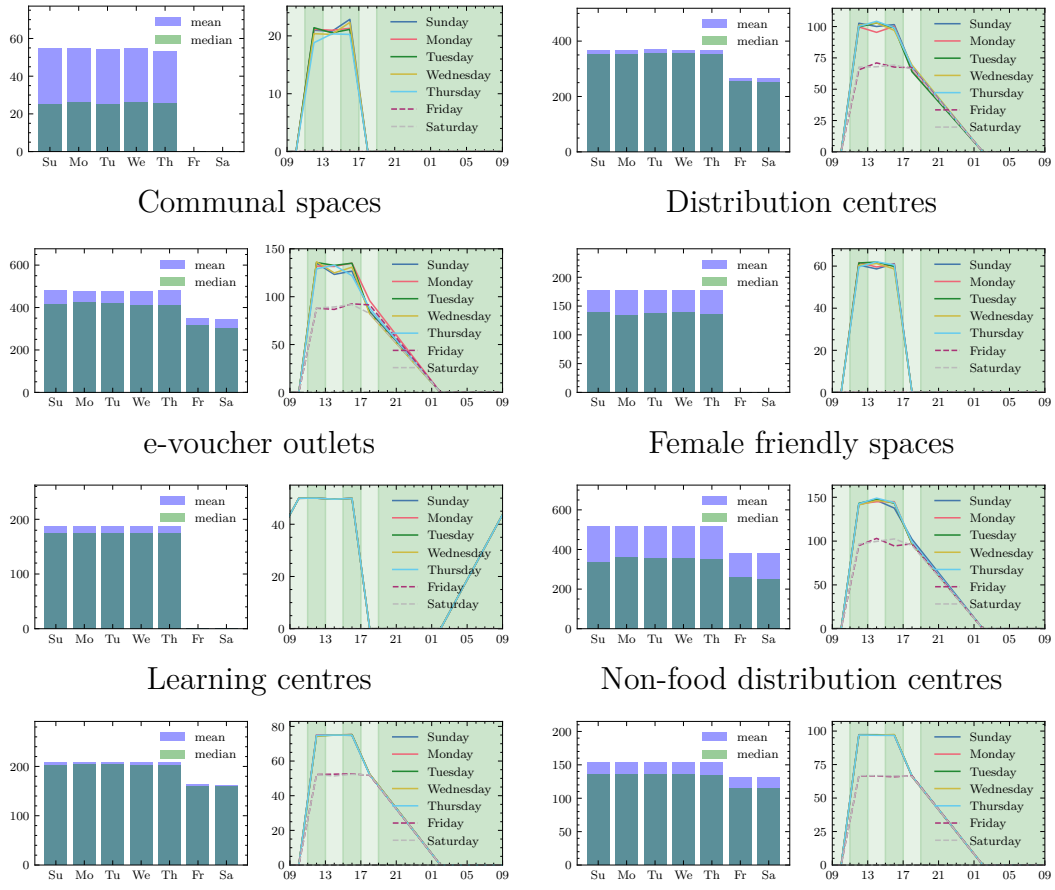


Figure 7.6: Left: The unique person attendance rates per day for the virtual venue. Right: The unique person attendance by time of day for the virtual venue. The green shading represents the discrete timestep bins of the simulation. JUNE can in general distinguish between days of the week however in JUNE-COX we only model differences between weekend and weekdays.

dividual being involved with *any* activity in a given time interval Δt is modelled through a Poisson process

$$P = 1 - \exp\left(- \sum_{i=1}^{N} \lambda_i(p) \Delta t \right). \tag{7.2.3}$$

If the agent participates in at least one of these activities, the specific activity i is selected according to

$$P_i = \frac{\lambda_i(p)}{\sum_{i=1}^{N} \lambda_i(p)}, \tag{7.2.4}$$

and the person is moved to the relevant location. If no activity is selected, the individual will stay in their shelter.

One of the outcomes of this exercise is condensed in Fig. 7.5, which shows the probabilities that men and women attend the different venues in the model as a function of their age.

The attendance probabilities were tuned to achieve the desired attendance rates reported in the questionaire Appendix C.3. These rates were chosen such that they represent an "average" day of any particular day of the week in the camp ignoring any changes of behaviour from religious or national events or annual variations in climate and weather.

It is important to stress that such census and demographic data is by default recorded by UNHCR and other non-governmental organisations (NGOs) operating in refugee and IDP settlements and it can be further supplemented or clarified by the survey described above or by interviews with settlement staff. This implies that it is a relatively straightforward exercise to apply our procedure outlined here to other settlements.

7.2.3 A Mixed-Method Approach

We have now set the stage to combine the information about the aggregate contact patterns with our highly-detailed model of interactions in a representative virtual population and to interrogate the model and extract detailed, survey informed, matrices. JUNE uses stochastic methods to simulate contacts between members of the virtual population which can be used to construct synthetic CMs. The random behaviour of the virtual population is encoded in repeatedly sampling the γ_{ij} from a Poisson distribution, $\tilde{\gamma}_{ij} \sim \mathcal{P}(\kappa_{ij})$ with the argument κ_{ij} distributed according to a normal distribution,

$$\kappa_{ij} \sim \mathcal{N}(\bar{\mu}, \sigma) \quad \text{with} \quad \mu = \frac{\Delta T}{T}\gamma_{ij} \quad \text{and} \quad \sigma = \frac{\Delta T}{T}\Delta\gamma_{ij}\,, \tag{7.2.5}$$

Venue, V	Characteristic time, T^L (hours)
Community centres	1.00 ± 0.02
Distribution centres	1.00 ± 0.05
e-voucher outlets	0.8 ± 0.1
Female friendly spaces	1.00 ± 0.02
Learning centres	1.5 ± 0.1
Non-food distribution centres	1.00 ± 0.05
Pump and latrines	0.7 ± 0.1
Religious centres	1.00 ± 0.01

Table 7.2: The characteristic time, T^L in hours for each venue reported in the survey. Value and error are determined using proportionate weighting between men and women using a median bootstrap method. Derived from Appendix C.3

with the γ_{ij} and their uncertainty ($\Delta\gamma_{ij}$) taken from the survey and re-scaled by the ratio of the typical time people attend a location, T (cf. Tab. 7.2) and the size of the emulation time-step in the model, ΔT. It is important to sample the argument κ_{ij} so that we encapsulate errors from the bootstrap methodology in the contact survey into our model. Finally we statistically round the individual instances $\tilde{\gamma}_{ij}$ to integer values, since the model agents can only contact a discrete number of other agents within the model with discrete characteristics. The resulting emulated set of $\tilde{\gamma}_{ij}$ are normalised such that they represent an individual agent's integer contacts per timestep. Averaging generates the $\hat{\gamma}_{ij}$ which can be directly compared with the γ_{ij} obtained from the survey.

In the simulation we aim to perform a virtual survey on the virtual population, as close as possible to the conditions in the real-world light-weight surveys. We sample individual behaviour over 28 virtual days to obtain individual $\tilde{\gamma}_{ij}$'s every time an agent attends a venue, the algorithm is outlined in Alg. 3.

The venues are filled according to the probabilities described above, Eq. (7.2.3), Eq. (7.2.4) and we "measure" the total raw contacts $\hat{t}_{ij}$ in the simulation. To further insure the correct total expected attendance time at the virtual venues compared with the real world, we proportionally close venues to approximate their possible

Alg. 3: The virtual survey. Loop over all venues and people and simulate P_{contacts} between i and j subgroups from survey. The contacts can then be clustered into arbitrary subgroups k, l. We allow for multiple contacts between the same people at venue L.

Input **Data:** $\hat{t}_{ij}^{L} = [0]_{kl}$
$\hat{t}_{kl}^{L}$, number of contacts between subgroup k and l at venue L,
$\hat{\eta}_{i}^{L}$, population of subgroup i at venue L,
P^{L}, list of agents at venue L,
T^{L}, total time,
ΔT, simulation time step,
$\tilde{\gamma}_{kl}^{L}$, Stochastic Poisson sampled contacts between subgroup k and l at venue L,
for $L \in$ Venues **do**
  for $P_x \in$ People @ L, P^{L} **do**
    $i = \text{subgroup}(P_x)$
    $T^{L} = T^{L} + \Delta T$
    $\hat{\eta}_{i}^{L} = \hat{\eta}_{i}^{L} + 1$ **for** $j \in L_{\text{subgroups}}$ **do**
      Generate $\tilde{\gamma}_{ij}^{L}$ **if** $\tilde{\gamma}_{ij}^{L} = 0$ **then**
        continue
      else
        Generate randomly a list of P_{contacts} of $\tilde{\gamma}_{ij}^{L}$ people at L in subgroup j not including P_x
      end
    end
    for $P_c \in P_{\text{contacts}}$ **do**
      $k = \text{subgroup}(P_x)$
      $l = \text{subgroup}(P_c)$
      $\hat{t}_{kl}^{L} = \hat{t}_{kl}^{L} + 1$
    end
  end
end
$\hat{t}_{kl}^{L} = \hat{t}_{kl}^{L} / T^{L}$

fractional opening times as compared with the chosen timesteps in the simulation. This procedure allows us to directly compare resulting matrices $\hat{t}_{ij}$, $\hat{\gamma}_{ij}$ and $\hat{c}_{ij}$ with their real-world counterparts t_{ij}, γ_{ij} and c_{ij} above. Even more, we are not constrained to the creation of virtual 2×2 contact matrices only, but can infer matrices for any sub-classification i and j that our simulation allows – in the results we present here, the i and j are age brackets of size 1 year. The final type(s) of contact matrix, PNCM_V and UNCM_V, $\hat{b}_{ij}$ and $\hat{\beta}_{ij}$, can also be calculated with a minor modification to the algorithm instead of generating a list of people p_j at the venue in contact with each person p_i, we allow "democratic/isotropic" contacts of all people:

$$\hat{\beta}_{ij} = \hat{\eta}_j - \delta_{ij} \,. \tag{7.2.6}$$

For each entry, ij, this represents the total contacts the $\hat{\eta}_i$ people with characteristics i at the venue have with the population of the venue in each subgroup. The Kronecker-δ corrects for "self-contacts".

7.3 Results

In this section we present the results of the contact matrices derived from our mixed-method approach. We begin by validating our method in the context of the UK where we compare our results against contact patterns directly collected by a traditional survey [135]. We recognise that the UK and Cox's Bazar are very different settings however, this validation step simply acts as a closure and sanity check in the context of a well understood and studied setting in which we can test our mixed method approach. Once our method has been validated, we present the matrices for the Cox's Bazar refugee settlement. Throughout, we use several key metrics to determine the similarity between any two sets of matrices:

1. Normalised Canberra distance, D_C [137]:

$$D_C(C, C') = \frac{1}{\text{Dim}(C) - Z} \sum_i \sum_j \frac{|C_{ij} - C'_{ij}|}{|C_{ij}| + |C'_{ij}|} \,, \tag{7.3.1}$$

where C and C' represent two contact matrices we wish to compare, Dim denotes the number of elements, $\mathrm{Dim}(C_{n\times m}) = n \cdot m$ and Z is the number of zero elements of the difference $(C_{ij} - C'_{ij})$.;

2. Q index as measure of assortativity [138],

$$Q = \frac{\mathrm{Tr}(C/\sum_{ij} C_{ij}) - 1}{\mathrm{Dim}(C) - 1} \,. \tag{7.3.2}$$

3. Dissimilarity index, I_s^2 [139]

$$I_s^2 = \frac{1}{2}\frac{\langle (X-Y)^2 \rangle_{F_c}}{\sigma_p^4} \,, \tag{7.3.3}$$

where σ_p is the standard deviation of the ages of the population and $\langle (X - Y)^2 \rangle_{F_c}$ represents the expectation age difference between contacts x and y of the function $F_c(x, y)$,

$$F_c(x,y) = \frac{f(x) C^L_{xy} f(y)}{\sum_x \sum_y f(x) C^L_{xy} f(y) \Delta x \Delta y} \,, \tag{7.3.4}$$

here Δx and Δy are the age bin sizes from the contact survey.

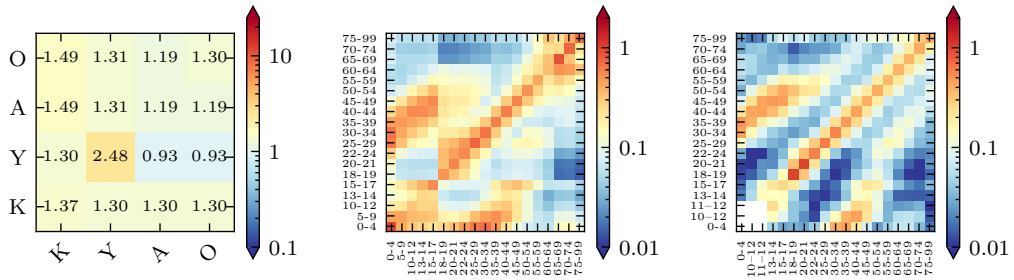
The normalised Canberra distance gives an estimation of the similarity between two matrices – approaching 0 when they are more similar and 1 when dissimilar. The remaining statistics measure the level of assortativity – the level of diagonal dominance and therefore the rate at which similar ages interact compared with dissimilar ages. The Q index ranges from 0 – homogeneous, proportionate mixing – to 1 – fully assortative. I_s^2 measures the deviation from perfect assortativity with a value of 0 when fully assortative and 1 for homogeneous interactions.

7.3.1 UK Validation

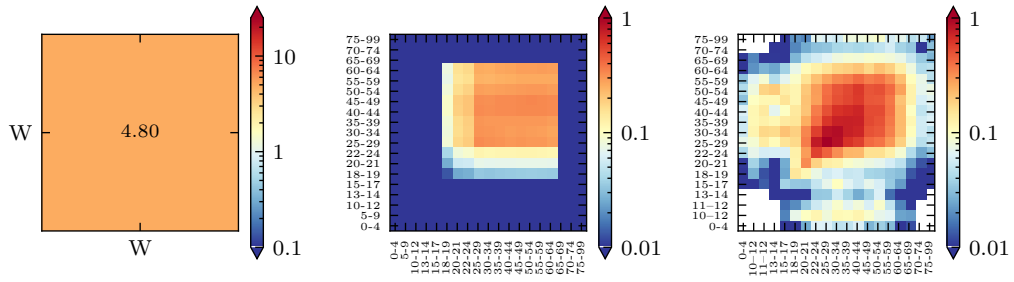
The first step of our virtual survey validation is to compare our results with that of real surveys conducted in far greater granularity. JUNE-UK has had extensive tuning for COVID-19 modelling in the UK [23, 128, 140]. As a proof-of-concept, we focus on

the most complex contact matrix – that of the household – and compare the contact matrices produced by the simulation with those from a traditional diary-based survey [135]. The input contact matrix is constructed from a combination of this data, the Office of National Statistics (ONS) census data of UK households [141, 142] and UK population demographics [143]. Since the UK census for household types distinguishes children (kids, K, <18 years old), young adults such as students or other dependent resident (Y, assumed 18-25 years old in JUNE-UK), adults (A, assumed 26-65 years old) and older adults (O, assumed >65 years old), we aggregated the granular contact matrix derived from the survey into a significantly coarser 4×4 matrix mapping the census categories. We also corrected for different household types to better incorporate the details of the venue-specific heterogeneities in their demographic composition. For more details on this procedure, see specifically Section 4 and Appendix C of the original description of the JUNE-UK modelling setup [23].

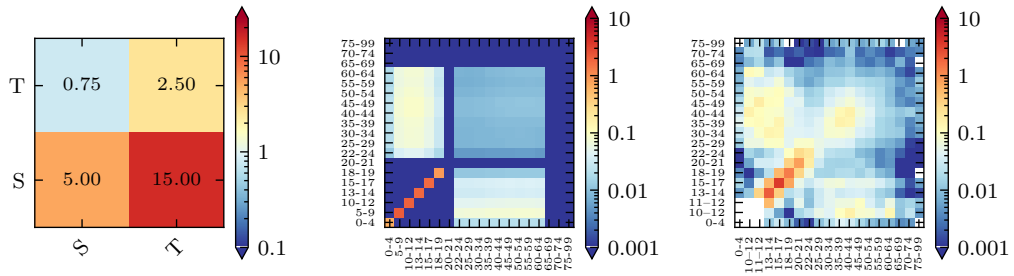
The results in Fig. 7.7 show the 4×4 input matrix derived from the aggregation process described above and a comparison of the output of the PNCM$_R$ $\hat{\Gamma}$ from the JUNE-UK model virtual contact survey with the results of the matrix C from the traditional survey. This provides a closure test ensuring that JUNE-UK returns realistic contact matrices from coarse aggregate matrices. This closure test is performed to explicitly demonstrate that the agent movement and contact tracing infrastructure in JUNE produces self consistent results provided the virtual settings are a realistic approximation of the real world and its social interactions. Using the setting of the UK for which a plethora of contact matrices and census information is available we can trust that the virtual survey in JUNE has a valid methodology. Clearly, our mixed-method approach is able to reproduce the broad structure of the real-world data – especially capturing the patterns of contacts between children and their parents represented in the off-diagonal structures. The original survey did not contain information on the contacts of younger children due to constraints on the data collection methodology; our method is able to fill this gap. This is shown in Fig. 7.7 where the JUNE matrices provide predictions of the child–child contacts for young



a) Household



b) Company



c) School

Figure 7.7: Left: The derived input interaction matrix, $\mathrm{UNCM}_{\mathbf{R}}$ for "Households", "Companies" and "Schools", where the labels "W" refers to "Workers", "S" students, "T", teachers. Center: The simulated age-binned $\mathrm{PNCM}_{\mathbf{R}}$ matrix with entries $\hat{C}_{ij}$ from JUNE-UK. Right: The BBC Pandemic project contact matrices, C, with entries c_{ij}.

children in the settings of households and schools.

To further validate our approach, we compare the Q, I_s^2 and D_C metrics of the two matrices. Tab. 7.3 shows that the first two metrics are in close agreement, with the overall Canberra distance being close to 0, thereby confirming the similarity of the matrices. Indeed, the difference between the measures of assortativity are comparable or better than those found in similar studies but which do not make

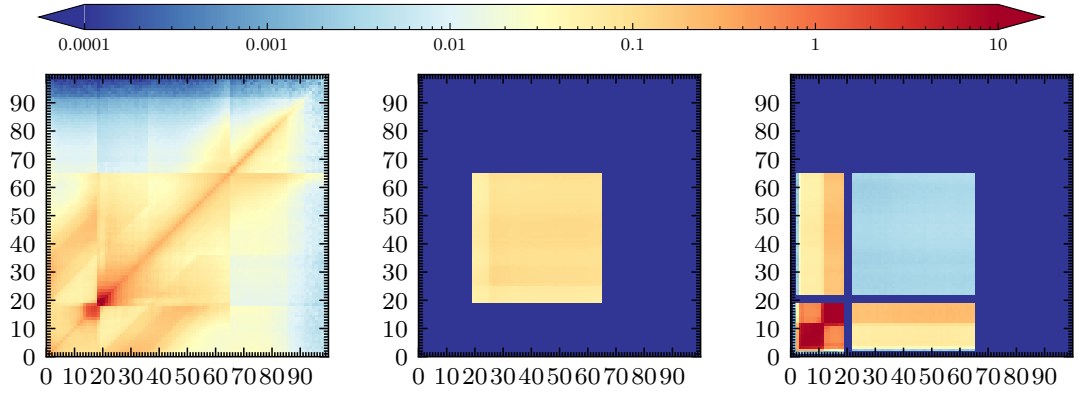
	Household			Company			School		
	Q	I_s^2	D_C	Q	I_s^2	D_C	Q	I_s^2	D_C
BBC Pandemic	0.14	0.36		0.026	0.31		0.13	0.14	
JUNE-UK PNCM$_{\mathbf{R}}$	0.12	0.30	0.32	0.0026	0.42	0.73	0.21	0.050	0.63

Table 7.3: Contact matrix statistics calculated for JUNE-UK and BBC Pandemic project reported for households, company and school mixing. These statistics are calculated for the UK demography reported by ONS in 2011 [143].

use of the guiding input aggregate matrix as we do here [127]. Given these strong findings, together with the visual and structural similarities of the matrices, we consider our mixed-method approach to be reasonably validated for application to settings in which intensive survey-based approaches to deriving contact patterns are not feasible. For real-world applications, we note that our methodology is clearly not exactly reproducing the original surveys; however, users will have to decide whether these errors are acceptable in comparison to having little or no knowledge about contact patterns, or making necessary assumptions about these patterns. We will discuss further in the conclusions of this work but this methodology is clearly subject to structure and rules applied in the simulation. It is also worth noting that the virtual agent behaviour of JUNE-UK is much better informed than those in JUNE-COX. This will become clear in the disparity between NCM, NCM$_{\mathbf{R}}$ and NCM$_{\mathbf{V}}$ type contact matrices. PNCM$_{\mathbf{V}}$ matrices presented in the Fig. 7.8 and PNCM$_{\mathbf{R}}$ matrices in Fig. 7.7 have the same general shape and scaling of features. In the case of JUNE-COX derived matrices NCM, NCM$_{\mathbf{R}}$ and NCM$_{\mathbf{V}}$ types are less similar.

7.3.2 Contact Matrices in Cox's Bazar Refugee Settlement

The derived contact matrices are presented at the end of this section (Section 7.3.2).



(a) Household: $Q = 4.9 \times 10^{-2}$, $I_s^2 = 0.16$

(b) Company: $Q = 5.8 \times 10^{-4}$, $I_s^2 = 0.45$

(c) School: $Q = 1.3 \times 10^{-2}$, $I_s^2 = 0.050$

Figure 7.8: The normalised venue contact matrices ($\text{PNCM}_{\mathbf{V}}$) by age as simulated in JUNE-UK.

The lightweight survey in the camp was conducted across the following venues: "community centres", "distribution centres", "e-voucher outlets" and "formal education centres". For the remaining two venues – "play groups" and "shelters" – we assume that everyone generally mixes with everyone else in that location given the assumed small groups of children who play together, as well as the dense shelter environments. Since certain shelters are shared between multiple families, we differentiate intra- and inter-family mixing with the latter being represented by the diagonal elements of the aggregate matrix (i.e. setting these to the number of contacts within each of the two families in the shelter and with the off diagonal elements set to the number of contacts between the families). As discussed in previous work [42], we set the number of contacts within the families or play groups to the average size of these respective groups assuming homogeneous mixing in these settings. In the case of the play groups we dis-aggregate the population into three age groups 3-6, 7-11 and 12-17 which mix homogeneously to emulate children typically interacting with children of similar age. We report the results for the $\text{UNCM}_{\mathbf{R}}$ γ_{ij} of the prior information and of the survey in Fig. 7.9 and Fig. 7.10. These CMs provide a closure test by comparing them to the UNCM $\hat{\mu}_{ij}$ results from performing a virtual survey in JUNE-COX with

the same coarse population categories. In the two figures we use the shorthand "T" and "S" for teachers and students in the learning centres and "H_x" for household x in a shared shelter.

Once we have determined the UNCM and confirmed that their stochastic uncertainties are within the uncertainties of the input interaction matrices, we can perform any custom binning for arbitrary group characteristics. Fig. 7.12 shows the final fully dis-aggregated (by age and venue) set of matrices for the Cox's Bazar refugee settlement based on the input contact matrices from the lightweight survey, combined with our highly-detailed agent-based model of the settlement. The combination of these two techniques leads to interesting consequences in the structure of the derived contact matrices. Contact rates from the light-weight survey provide the baseline coarse social interaction patterns between broad subgroups at a given venue. Whereas, the agent-based model embeds the dynamics from data on the social behavior of individuals, connecting many independent venues within the model. In particular, we see bands due mainly to 11-18 year olds for two reasons. Firstly, many behavioural patterns are defined differently for adults and children leading to attendance differences at 18. Secondly, at 11 years of age men are permitted to attend the religious centres. Due to the high rate of attendance observed at the religious centres, there is a drop in attendance at other non-religious centre venues of this age group relative to other age groups.

In Fig. 7.12, we can clearly see the effects of the different age groups as a consequence of the underlying social dynamics. For example, we observe large differences in the number of contacts between all age groups with adults in the community centres relative to the distribution centres, with substructures based on the age profile of children attending these locations shown through the higher number of contacts in younger age brackets. In addition, the learning centre matrices show a clear mix of contacts between children in their mixed classes and their teachers – this matrix also encodes information on the enrollment rate of children in the education system, with lower enrollment rates as the age of children increase. Finally, the detailed

information available on household and shelter composition appears in the shelter contact matrix which contains a number interesting features. We reconstruct a strong leading diagonal which represents persons of similar ages living together; siblings, parents and grandparents of similar ages the width of the band reflects spousal age gaps and minimal age gaps between consecutive siblings. Using more detailed information about the average age of parents at the birth of their first child we also develop off-diagonal structure in the upper left and lower right quadrants. There is an almost linear structure corresponding to children and parents interacting and aging together. This structure then tapers off indicating interactions in multi-generational households before many children would leave home at around 18.

A simpler approach is to just assume that everyone contacts everyone else in these dense settings in the absence of other information – we also present the results for the corresponding UNCM_V in Fig. 7.13. However, clearly there is a significant loss of information in doing this, in comparison to the mixed-method approach, as can be seen in the absence of structural detail in many of the UNCM_V matrices.

Population normalised matrices can be calculated from the user normalised matrices with a simple re-scaling as described above. We present these for completeness and the varied utility of each normalisation in different model types in, Fig. 7.14, Fig. 7.15, Fig. 7.16.

Here we present the contact matrices derived from JUNE-COX.[1] Each of the contact matrix elements derived from JUNE come with an associated standard error, these errors are typically very small due to the long run time (28 simulated days) and the small variance in overall attendance rates of difference demographic groups at each of the venues.

[1]We interpret contact matrix Δ_{ij} such that person i contacts person j and graphically as subgroup on x-axis contacts subgroup on y-axis.

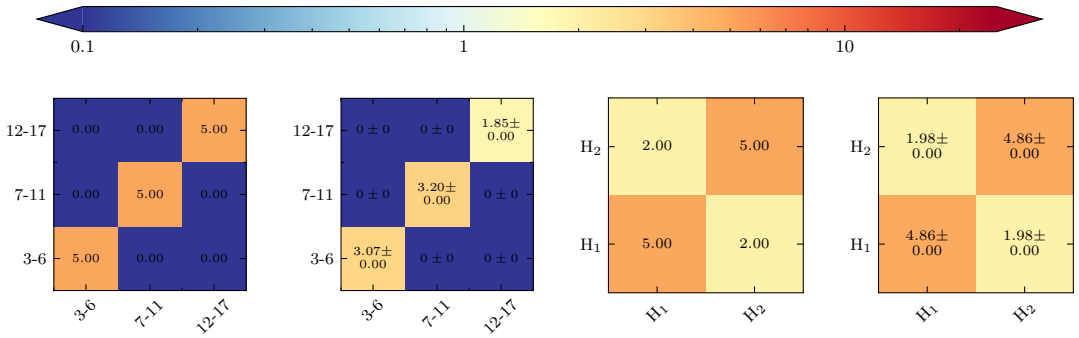
Presentation of Contact Matrices

The first set of matrices reported here are those which provide the closure test where we directly compare the $\text{UNCM}_{\mathbf{R}}$ contact matrices from the real world survey with the virtual survey with the survey subgroups:

1. Non-survey derived $\text{UNCM}_{\mathbf{R}}$ matrices (cf. Fig. 7.9),
2. Survey derived $\text{UNCM}_{\mathbf{R}}$ matrices (cf. Fig. 7.10).

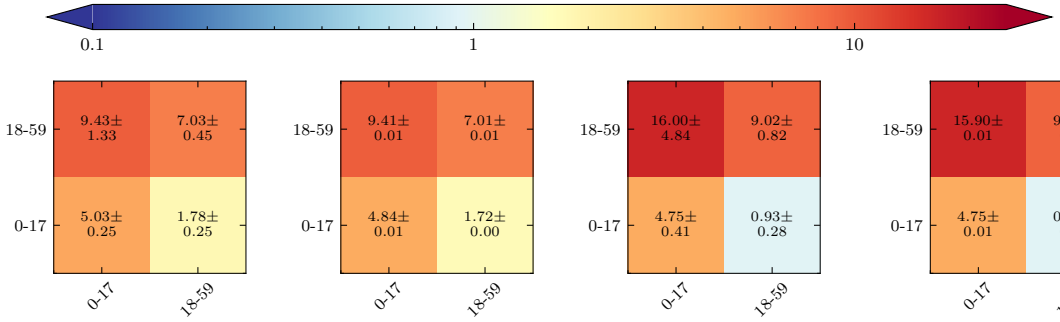
The second set of matrices reported here are fully age disaggregated are constructed from the virtual survey. We report:

1. UNCM (cf. Fig. 7.11),
2. $\text{UNCM}_{\mathbf{R}}$ (cf. Fig. 7.12),
3. $\text{UNCM}_{\mathbf{V}}$ (cf. Fig. 7.13),
4. PNCM (cf. Fig. 7.14),
5. $\text{PNCM}_{\mathbf{R}}$ (cf. Fig. 7.15),
6. $\text{PNCM}_{\mathbf{V}}$ (cf. Fig. 7.16).

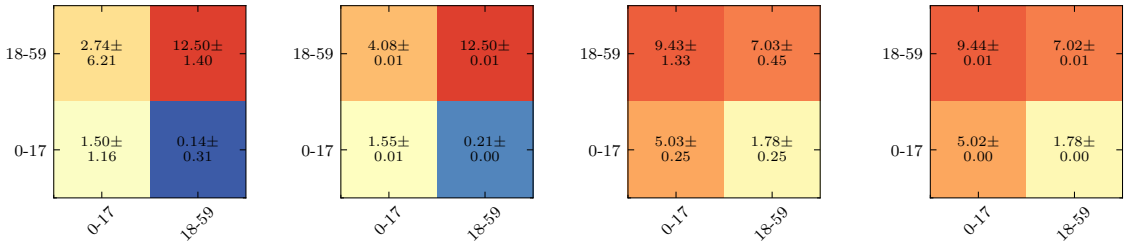


Play groups: $D_C = 3.0 \times 10^{-1}$ Shelters: $D_C = 9.2 \times 10^{-3}$

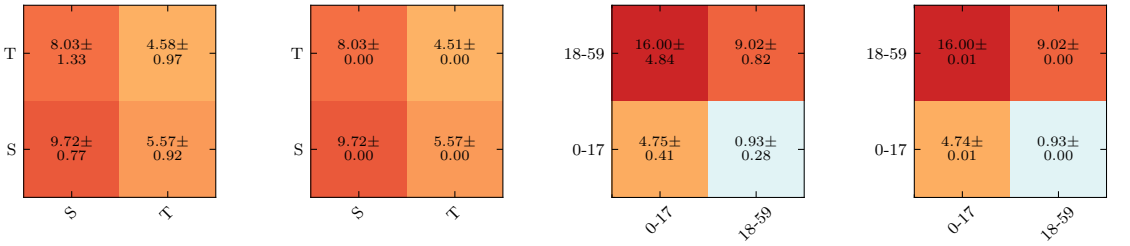
Figure 7.9: Left: Two pairs of $\text{UNCM}_{\mathbf{R}}$ for the virtual venues determined prior to the light-weight survey. Right: The JUNE-COX virtual survey in the same coarse population bins. Captions reporting the Canberra distance between them.



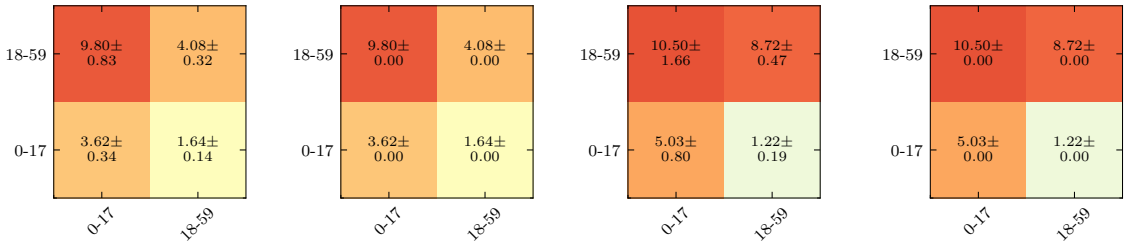
Community centres: $D_C = 7.6 \times 10^{-2}$ Distribution centres: $D_C = 2.0 \times 10^{-1}$



e-voucher outlets: $D_C = 2.0 \times 10^{-1}$ Female friendly spaces: $D_C = 1.9 \times 10^{-1}$

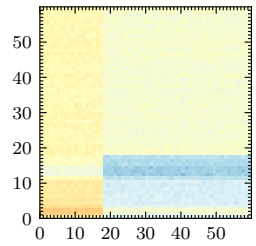


Learning centres: $D_C = 2.7 \times 10^{-1}$ n.f. distribution centres: $D_C = 2.0 \times 10^{-1}$

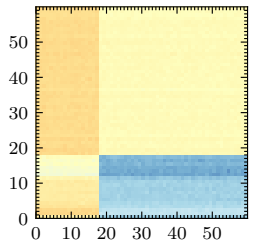


Pumps and latrines: $D_C = 2.2 \times 10^{-1}$ Religious centres: $D_C = 1.5 \times 10^{-1}$

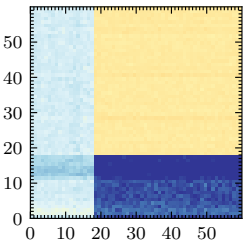
Figure 7.10: Left: The $\text{UNCM}_{\mathbf{R}}$ from the contact survey data. Right: JUNE-COX virtual survey UNCM (Right), with the relative Canberra distances. We set "Community centres" and "Distribution centres" identical to "Female friendly spaces" and "Non-food distribution centres", respectively.



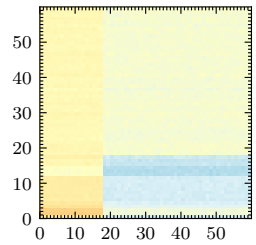
(a) Community centres:
$Q = 1.5 \times 10^{-3}$
$I_s^2 = 0.48$



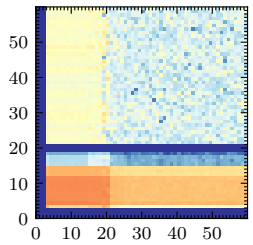
(b) Distribution centres:
$Q = 1.4 \times 10^{-3}$
$I_s^2 = 0.49$



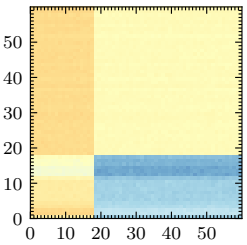
(c) e-voucher outlets:
$Q = 2.5 \times 10^{-3}$
$I_s^2 = 0.32$



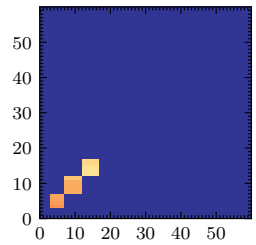
(d) Female friendly spaces:
$Q = 1.4 \times 10^{-3}$
$I_s^2 = 0.43$



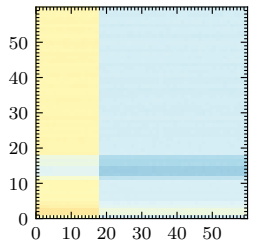
(e) Learning centres:
$Q = -1.8 \times 10^{-3}$
$I_s^2 = 0.27$



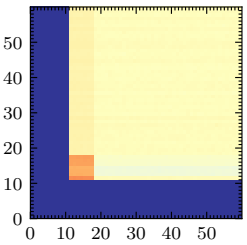
(f) n.f. distribution:
$Q = 1.5 \times 10^{-3}$
$I_s^2 = 0.49$



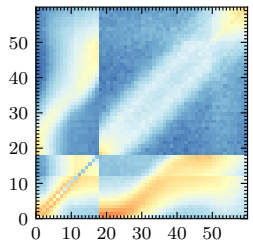
(g) Play groups:
$Q = 2.0 \times 10^{-2}$
$I_s^2 = 0.056$



(h) Pumps and latrines:
$Q = -1.2 \times 10^{-4}$
$I_s^2 = 0.49$

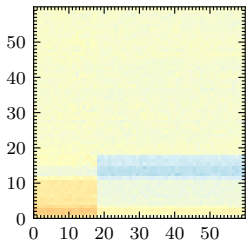


(i) Religious centres:
$Q = 2.0 \times 10^{-3}$
$I_s^2 = 0.41$

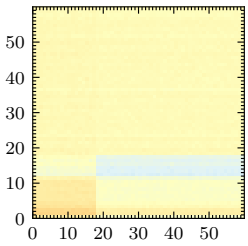


(j) Shelters:
$Q = 7.7 \times 10^{-3}$
$I_s^2 = 0.38$

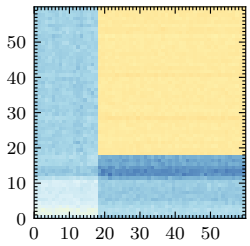
Figure 7.11: The normalised contact matrices (UNCM) by age as simulated in JUNE-COX. Note that the data inputs in (g) and (j) stem from a previous survey.



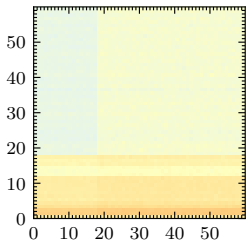
(a) Community centres:
$Q = 1.5 \times 10^{-3}$
$I_s^2 = 0.48$



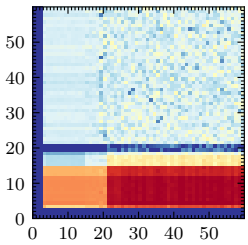
(b) Distribution centres:
$Q = 7.7 \times 10^{-4}$
$I_s^2 = 0.49$



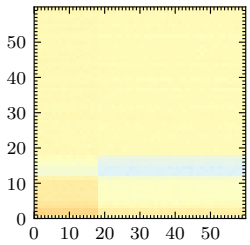
(c) e-voucher outlets:
$Q = 3.3 \times 10^{-3}$
$I_s^2 = 0.31$



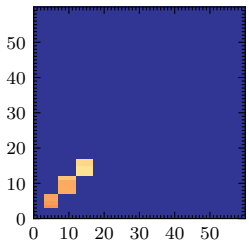
(d) Female friendly spaces:
$Q = 3.2 \times 10^{-4}$
$I_s^2 = 0.50$



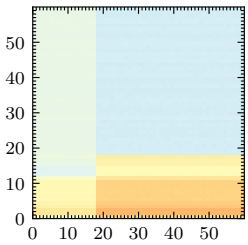
(e) Learning centres:
$Q = -1.8 \times 10^{-3}$
$I_s^2 = 0.88$



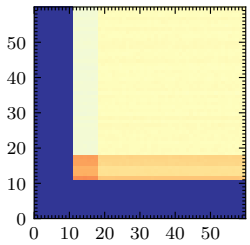
(f) n.f. distribution:
$Q = 8.6 \times 10^{-4}$
$I_s^2 = 0.49$



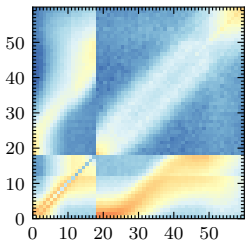
(g) Play groups:
$Q = 2.0 \times 10^{-2}$
$I_s^2 = 0.056$



(h) Pumps and latrines:
$Q = -2.3 \times 10^{-3}$
$I_s^2 = 0.60$

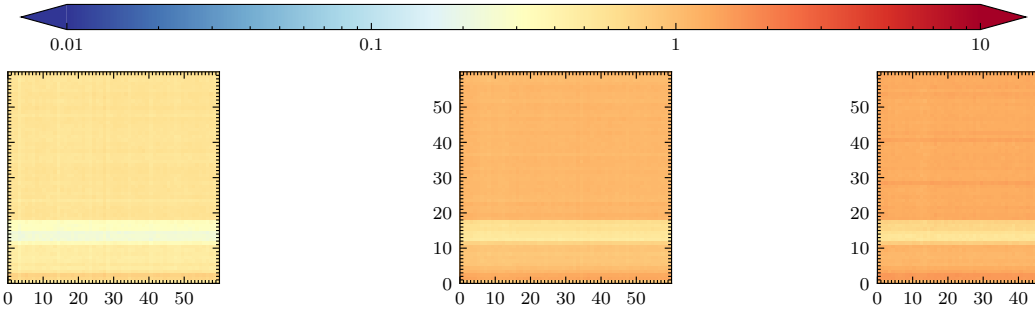


(i) Religious centres:
$Q = 1.9 \times 10^{-3}$
$I_s^2 = 0.45$



(j) Shelters:
$Q = 7.4 \times 10^{-3}$
$I_s^2 = 0.38$

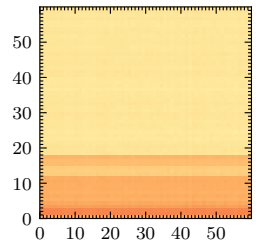
Figure 7.12: The reciprocal normalised contact matrices ($\text{UNCM}_{\mathbf{R}}$) by age as simulated in June-Cox. Note that the data inputs in (g) and (j) stem from a previous survey.



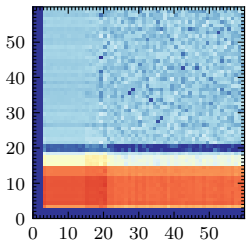
(a) Community centres:
$Q = -9.2 \times 10^{-6}$
$I_s^2 = 0.50$

(b) Distribution centres:
$Q = 1.6 \times 10^{-5}$
$I_s^2 = 0.50$

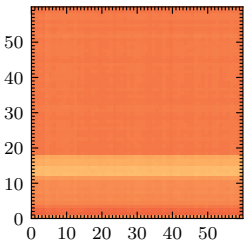
(c) e-voucher outlets:
$Q = -2.6 \times 10^{-5}$
$I_s^2 = 0.50$



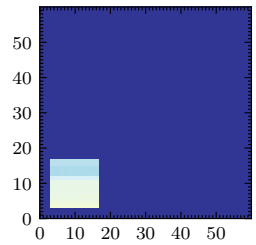
(d) Female friendly spaces:
$Q = 1.5 \times 10^{-5}$
$I_s^2 = 0.50$



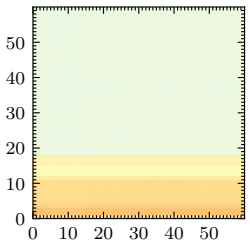
(e) Learning centres:
$Q = 3.9 \times 10^{-4}$
$I_s^2 = 0.28$



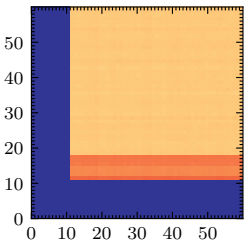
(f) n.f. distribution:
$Q = 1.9 \times 10^{-5}$
$I_s^2 = 0.50$



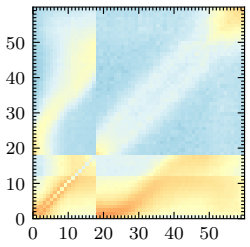
(g) Play groups:
$Q = 3.2 \times 10^{-5}$
$I_s^2 = 0.47$



(h) Pumps and latrines:
$Q = 4.9 \times 10^{-5}$
$I_s^2 = 0.48$

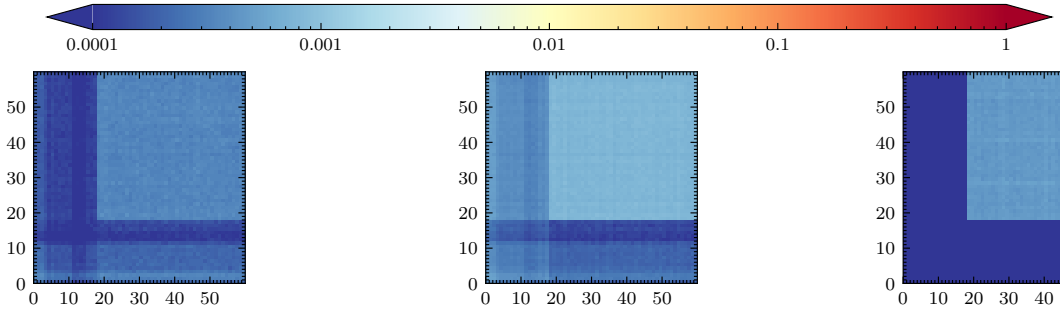


(i) Religious centres:
$Q = 1.1 \times 10^{-5}$
$I_s^2 = 0.52$



(j) Shelters:
$Q = 4.9 \times 10^{-3}$
$I_s^2 = 0.41$

Figure 7.13: The normalised venue contact matrices ($\text{UNCM}_{\mathbf{V}}$) by age as simulated in JUNE-COX. Note that the data inputs in (g) and (j) stem from a previous survey.



(a) Community centres:
$Q = 1.5 \times 10^{-3}$
$I_s^2 = 0.43$

(b) Distribution centres:
$Q = 1.4 \times 10^{-3}$
$I_s^2 = 0.39$

(c) e-voucher outlets:
$Q = 2.5 \times 10^{-3}$
$I_s^2 = 0.28$

(d) Female friendly spaces:
$Q = 1.4 \times 10^{-3}$
$I_s^2 = 0.44$

(e) Learning centres:
$Q = -1.3 \times 10^{-5}$
$I_s^2 = 0.16$

(f) n.f. distribution:
$Q = 1.5 \times 10^{-3}$
$I_s^2 = 0.40$

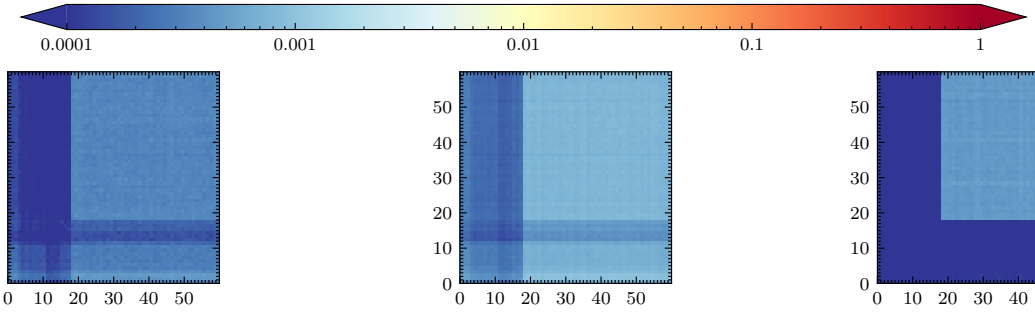
(g) Play groups:
$Q = 2.0 \times 10^{-2}$
$I_s^2 = 5.4 \times 10^{-2}$

(h) Pumps and latrines:
$Q = -1.2 \times 10^{-4}$
$I_s^2 = 0.50$

(i) Religious centres:
$Q = 2.0 \times 10^{-3}$
$I_s^2 = 0.41$

(j) Shelters:
$Q = 7.7 \times 10^{-3}$
$I_s^2 = 0.38$

Figure 7.14: The normalised venue contact matrices (PNCM) by age as simulated in JUNE-COX. Note that the data inputs in (g) and (j) stem from a previous survey.



(a) Community centres:
$Q = 1.5 \times 10^{-3}$
$I_s^2 = 0.52$

(b) Distribution centres:
$Q = 7.7 \times 10^{-4}$
$I_s^2 = 0.46$

(c) e-voucher outlets:
$Q = 3.3 \times 10^{-3}$
$I_s^2 = 0.29$

(d) Female friendly spaces:
$Q = 3.2 \times 10^{-4}$
$I_s^2 = 0.52$

(e) Learning centres:
$Q = -1.8 \times 10^{-3}$
$I_s^2 = 0.21$

(f) n.f. distribution:
$Q = 8.6 \times 10^{-4}$
$I_s^2 = 0.47$

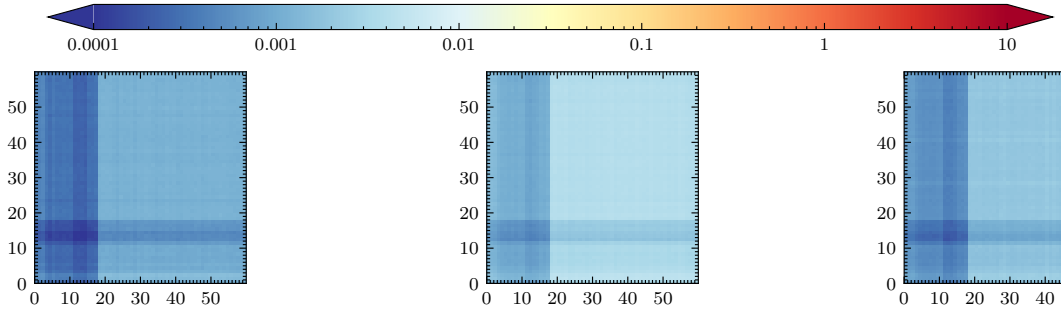
(g) Play groups:
$Q = 2.0 \times 10^{-2}$
$I_s^2 = 0.054$

(h) Pumps and latrines:
$Q = -2.2 \times 10^{-3}$
$I_s^2 = 0.6$

(i) Religious centres:
$Q = 1.9 \times 10^{-3}$
$I_s^2 = 0.47$

(j) Shelters:
$Q = 7.4 \times 10^{-3}$
$I_s^2 = 0.39$

Figure 7.15: The population normalised contact matrices ($\text{PNCM}_{\mathbf{R}}$) by age as simulated in JUNE-COX. Note that the data inputs in (g) and (j) stem from a previous survey.



(a) Community centres:
$Q = -9.2 \times 10^{-6}$
$I_s^2 = 0.46$

(b) Distribution centres:
$Q = 1.6 \times 10^{-5}$
$I_s^2 = 0.46$

(c) e-voucher outlets:
$Q = -2.6 \times 10^{-5}$
$I_s^2 = 0.45$

(d) Female friendly spaces:
$Q = 1.5 \times 10^{-5}$
$I_s^2 = 0.52$

(e) Learning centres:
$Q = 3.9 \times 10^{-4}$
$I_s^2 = 0.13$

(f) n.f. distribution:
$Q = 1.9 \times 10^{-5}$
$I_s^2 = 0.47$

(g) Play groups:
$Q = 3.2 \times 10^{-5}$
$I_s^2 = 0.44$

(h) Pumps and latrines:
$Q = 4.9 \times 10^{-5}$
$I_s^2 = 0.48$

(i) Religious centres:
$Q = -1.1 \times 10^{-5}$
$I_s^2 = 0.52$

(j) Shelters:
$Q = 4.9 \times 10^{-3}$
$I_s^2 = 0.42$

Figure 7.16: The normalised venue contact matrices ($\text{PNCM}_{\mathbf{V}}$) by age as simulated in JUNE-COX. Note that the data inputs in (g) and (j) stem from a previous survey.

7.4 Discussion

The matrices presented in this work are among the first contact matrices derived for a refugee settlement. While not collected using traditional survey methods, we use a mixed-method approach for their calculation, which presents a new way to collect contact data. This is particularly useful in settings, such as in refugee settlements, in which data collection can present many challenges and therefore needs to be lightweight and integrated in to existing data collection regimes and programming.

We are able to perform closure tests on the contact matrices we derive and show that they clearly demonstrate great potential for a lightweight survey and an agent-based model to provide deeper insights into social environments when combined together. The survey and JUNE-COX derived contact matrices are initially validated by a comparison of their Canberra distances and their elements over the survey subgroups ij. These values are found to be very close to zero with the exception of the e-voucher outlets in which child – child contacts are higher in JUNE-COX than reality. This discrepancy can be explained by considering that the survey has a high uncertainty in the expected child – child contacts, an error in which JUNE-COX incorporates into the contact tracking algorithm. That is the large error in the number of contacts causes the children to contact a large range of people despite the low probability of attendance of children reported at the e-voucher venues. The children agents that do attend an e-voucher venue will sample the number of contacts they should make from a broader distribution due to this larger error from the survey. Further validation is performed with JUNE-UK derived matrices on age-disaggreagated contact matrices in which we are able to use other statistics such as I_S^2 and Q. These matrices were found to be in good agreement with other more intensive contact surveys. This validation ensures that the combination of coarse input contact matrices and the attendance rates responsible for agent dynamics yield representative contact patterns over all ages.

In the case of refugee settlements, the derived contact matrices can be used to

understand the social contact patterns using data already collected regularly by international organisations such as UNHCR, while being supplemented by data which can easily be collected by enumerators in a resource-efficient way. The highly-detailed matrices derived for the Cox's Bazar settlement demonstrate clear inter-age mixing patters which are crucial inputs to other epidemic models to represent realistic social mixing patterns. In particular, clear features are present in the matrices due to differing attendance rates and household compositions.

From the technical perspective, there are several further considerations and limitations to this methodology that become apparent when analysing the full age-disaggregated contact matrices (see Section 7.3.2). These pertain to the way in which the data is collected and the model is constructed and can be used as ways to diagnose the performance of the method:

1. **Subgroup classification:**

   Subgroup classification refers to the broad definition of subgroups defined in the model. Throughout June-Cox and June-Uk we define "Adults", "Children", "Teachers", "Workers" etc. which all have unique parameters and rules governing their behaviour. Subgroups defined by age can lead to strong banding artifacts in the contact matrices. These effects can be mitigated by blurring the age cut-off with some finite probability – e.g. that a child of 17 may behave like an adult. This mitigation should only be implemented in situations in which we are certain that there should not be a discontinuity in behaviours in the real world. For example, only over 11 year old men are permitted to attend the religious centres and hence we expect a sharp cut off in the contact matrices whereas in many other venues we expect a gradual shift in behaviour as children move into adolescence and then adulthood. This can be a positive feature of the model – i.e. that the model represents the behavioural and movements patterns correctly and forces agents to make a choice between activities they perform as they would in real life – however, this relies on reasonable behavioural data, insights and assumptions. This is demonstrated most clearly

in the *shelters* contact matrices in which the household clustering places adults and children differently based on fixed rules derived from survey and census data.

2. **Virtual venue demography:**

   The dynamics of virtual spaces in the simulation are dictated by the probabilistic attendance rates (see Fig. 7.5) and age cut offs. The attendance rates are a function of age, sex, time and venue which leads to different demographies across the virtual spaces and therefore different social mixing behaviours. Again, due to the nature of the simulation in which we have strict probabilistic rules which determine the attendance of different subgroups (children, adults, age or sex etc.), we can get strong divisions between groupings. This is shown by the discontinuities in the heat-map representation of the contact matrices. In particular, only men over the age of 11 are permitted to attend the religious centres leading to discontinuities in the religious centre contact matrices. In JUNE-UK, there is no simulation of parent-teacher interactions at school that might occur during pick up or drop off times and the virtual school setting is strictly modelling student-teacher interactions where any teacher-teacher interactions would be restricted to the classroom setting. Further, no children attend any work place settings and agents can only be employed or attend a work place venue between the ages of 18-65. The contact matrices produced from JUNE-UK therefore lack certain features shown by the BBC Pandemic project. However, this is a problem all such approaches that rely on an imperfect virtual representation of reality can experience.

3. **Virtual world rules and behaviour patterns:**

   The combination of the above points leads to complex inter-connected behaviours across the simulation. Considering the behaviour of coarser subgroups across all venues we see more general behaviours emerge; for instance, children are less likely to attend any virtual venue than adults and men are more likely

to attend any venue than women due to the attendance at religious centres which increases the overall rate of men not staying in the shelters compared to women, leading to an asymmetry in the shelter contact matrix. An 11-18 year old is more likely to see a 6-11 year old than the converse. A 6-11 year old is more likely to be home than a 11-18 year old therefore on average in any timestep a 6-11 year old will not contact an 11-18 year old in shelters, but when the 11-18 year old is home they will likely contact the 6-11 year old. The normalisation of contacts by users (or population) and contact duration (as done throughout) makes this effect visible. There are other instances, such as the community centres, in which we see a banding effect which is an induced artifact from the movement criterion of the agents in the model. The high attendance rate expected of 11+ men leads to a reduction in attendance of this group across all other venues and many of the contact matrices show a banding effect between 11 – 18 due to this behaviour.

Given the level of detail contained within the model-derived contact matrices, they have the ability to reveal potential short-comings in both the survey setup as well as the modelling of the virtual world, as they reflect how sophisticated and well understood each venue type is. This means that the amount of resources needed to be expended on collecting more data on certain locations can be estimated in order to improve certain matrices. These can be traded-off against the resources available and the relative expected gain from their expenditure. In this work, we validated our contact tracker in two very different models, June-Uk and June-Cox. In the former, we demonstrated that the $\mathrm{NCM}_{\mathbf{R}}$ agree well with data collected using traditional methods (cf. Fig. 7.7, Tab. 7.3). In the latter, $\mathrm{NCM}_{\mathbf{V}}$ type contact patterns are not available as our extracted contact matrices used coarse survey information on venue attendance to inform the simulation of contact patterns there, with the notable exception of the shelters, which are relatively precisely captured by the census data. Our mixed-method approach allows us to partially compensate for the gaps in detailed understanding of demographic structures at the lesser-known

venues.

7.5 Conclusion

In this work we demonstrate the complementary power of a lightweight contact survey, approximate details about venues and their attendance rates by different demographic groups and an agent-based model to generate detailed social contact matrices. In the case of the Cox's Bazar refugee settlement, we use an existing model of the settlement developed using the JUNE framework to perform a virtual contact survey, which is informed by the highly aggregate real world survey, to produce more granular contact matrices which can be further interrogated. Our constructed contact matrices will provide an important input to future disease spread modelling or social dynamic studies in the settlement and provide a baseline which can be translated to other settlements as well. Further, our method can easily be adapted to other settings for which detailed contact matrices are not available, thereby enabling the use of disease models in contexts where previously large assumptions would have had to have been made about contact patterns. Contact matrices form the backbone of many disease models and so calculating them at a global scale, with the specific inclusion of those groups who are often most vulnerable to disease spread, is essential [118].

Chapter 8

Conclusions

In this thesis, we apply a range of data intensive and machine learning methods to a range of research domains. We see that these tools are exceedingly powerful in dissecting extremely complex, large multi-dimensional datasets and distilling the information they represent into something useful and comprehensible.

In particular, we start by using SHERPA as a Standard Model Monte Carlo event generator in which we can determine present and future constraints on the upper bound of the Yukawa charm coupling. We configure SHERPA to produce a selection of signal and background processes in three key channels which can be distinguished by the isolated lepton count in the final state, namely: Vector Boson Fusion, W Higgs-strahlung and Z Higgs-strahlung Higgs production channels. The signal processes are defined by those in which the Higgs boson decays into two charm quarks directly. We examine the jet structure of the boosted fatjet produced by Higgs' decay into two charm quarks and subsequent decays. The structure of this fatjet and its subjets are scrutinised using a range of observables which we use to build a cut-flow first heuristically and then with cuts informed by neural networks. Two multivariate neural network architectures are used which take a selection of data input types. One network aims to distinguish processes with Higgs fatjets from those without and the second aims to distinguish Higgs to charm decays from Higgs to bottom decays. The first network takes inputs of jet observables, jet images and particle flow

information and the second takes displaced vertex observables and vertex particle flow information. This cut-flow eliminates several orders of magnitude of the background processes compared with the signal processes and the remaining surviving events are used to produce a selection of histograms of observables. These histograms are used to perform a statistical analysis under the κ-framework using the $\mathrm{CL_s}$ method. The κ-framework provides us with a direct way to relate the signal strength modifier to a value which directly modifies the Yukawa coupling. The neural network based cuts eliminate enough background so we can extract the Yukawa charm coupling. Thus, we can determine confidence limits on the Yukawa coupling. The derived coupling confidence limits are competitive to the current limits placed by experiments. Further, it is shown that novel choices in the profiling histogram could provide better constraints on couplings profiled under the κ-framework than the typical choice of invariant jet mass. We produce a set of competitive limits and projections on the Yukawa coupling compared with those produced by ATLAS and CMS.

Next we turn to the unsupervised methods of principal component analysis and Bayesian change point detection to understand underlying distributions of various instrumentation measurements at Hartlepool power station. Changes in the behaviour of the measurements could be indicative of faults or unexpected load drops. We examined two types of regions: *monitor* those known to precede faults and load drops versus, *healthy* those understood to experience nominal reactor behaviour. We employ principal component analysis and Bayesian online change point detection as tools in anomaly detection, due to their efficiency and ready interpretability as opposed to neural networks which are far less interpretative. The ability to understand the analysis that the modelling tool performs is paramount, particularly at a nuclear power station. Although the labeling procedure used to classify between the *healthy* and *monitor* regions is naive, the tools explored here demonstrated a clear classifying ability. These methods were packaged into a demonstration dashboard for further investigation by the engineers and scientists at Hartlepool power station.

Analysis tools which provide further insight into unexpected or anomalous behaviour can only improve the efficiency and safe operation of the power plant. This allows for more efficient replacement of faulty instrumentation and more intuitive operational tools for engineer interpretation.

Lastly, we apply rigorous methods and tools from physics and epidemiology to the social sciences and in particular we focus on using large agent based models to extract social mixing contact matrices. These contact matrices are particularly important in epidemiological models to understand how a disease will spread through a population in a healthcare emergency. In particular, we focus on Cox's Bazar, a refugee camp containing a population of $\sim 600,000$ vulnerable people. We develop algorithms in JUNE – an agent based epidemiological simulation tool – to track agents over time and their social interactions within the model at various social settings. At each timestep in the simulation, we conduct a virtual contact survey at all venues which can be used to determine social mixing matrices of various types. The algorithms and synthetic contact matrices are validated in JUNE-UK, a version of the model in which social settings and parameters of the model have been carefully tuned using census data and COVID-19 modelling data. The produced synthetic contact matrices can be directly compared with real-world survey derived contact matrices. With this satisfactory validation procedure we produce a selection of contact matrices for Cox's Bazar, which will be utilised in future work to better aid crisis response and refugee camp development.

While this thesis is broad in scope, we have only scratched the surface of the complexity, applicability and functionality of machine learning and data intensive methods that are available. As these tools continue to develop and the efficiency of processing larger datasets continues to improve, we will find that they will only become more powerful allies in data wrangling, comprehension and processing for years to come. Machine learning and data intensive tools encourage promising advancements across all fields of human endeavour, from understanding the smallest subatomic particles to ensuring the safe operation of carbon neutral power for the future or indeed,

tackling complex humanitarian challenges of our time.

Appendix A

The Standard Model

The Standard Model of particle physics is the one of the most precise, accurate and most *complete* theories in science. The theory explains the fundamental interactions of known matter with three of the four forces of nature, namely; the strong, weak and electromagnetic forces. The SM has undergone repeated tests over many decades providing theoretical predictions for particles and their masses, couplings and interactions. One of the many great successes of the SM is the most recent discovery of the Higgs boson in 2012 at the Large Hadron Collider with the ATLAS and CMS experiments [52, 53], this new scalar particle is consistent with the long theorised boson by Peter Higgs' (and others) in 1964 [144–148]. Despite the apparent great successes of the SM there still exist phenomena that are currently unexplained:

- Neutrino oscillations: What generates the Neutrino masses?
- Cosmological measurements: What is Dark Matter and Dark Energy?
- The CP problem: Why is there matter-antimatter asymmetry?
- Flavour problem: Why do the mass scales of the fermions span many orders of magnitude?
- Gravity: How does the fourth force of nature fit in?

Many researchers spend large portions of time developing beyond the standard model (BSM) theories to attempt to answer these pressing questions and issues and as of today no particles or forces from beyond the SM have been found at the LHC. As no new particles have shown up in the last decade we instead find ourselves in an era of precision measurements. That is, we aim measure couplings and the free parameters of the SM more and more precisely so that we might better understand the parts of the SM which have tensions against reality – where BSM physics might be lurking. One such choice of free parameters of the SM is shown in Tab. A.1.

Name	symbol	#	Comments
Quark Yukawa couplings:	y_u, y_d, y_s, y_c, y_b, y_t	6	} or masses
Lepton Yukawa couplings:	y_e, y_μ, y_τ	3	} or masses
CKM mixing angles:	θ_{12}, θ_{13}, θ_{23}	3	
CKM CP violating phase:	δ	1	
Gauge couplings:	g, g_Y, g_s	3	} or m_Z, m_W, α_s
QCD vacuum angle:	θ_{QCD}	1	
Higgs vacuum expectation value:	ν	1	
Higgs mass:	m_h	1	
		19	

Table A.1: One such parametrisation of the 19 free parameters of the SM.

The SM is formulated into a quantum field theory (QFT) in which the particle content is summarised in Tab. A.2, that is, each of these particles are treated as excitations in quantum fields in which their interactions are determined by the following gauge structure,

$$\mathrm{SU(3)}_c \times \mathrm{SU(2)}_\mathrm{L} \times \mathrm{U(1)}_\mathrm{Y} \, .$$

The individual gauge groups $\mathrm{SU(3)}_c$, $\mathrm{SU(2)}_\mathrm{L}$ and $\mathrm{U(1)}_\mathrm{Y}$ each give rise to vector gauge bosons that mediate a force of nature. $\mathrm{SU(3)}_c$ is responsible for the strong force and its interactions – Quantum Chromodynamics (QCD) – where c refers to colour charge with values *red*, *green*, *blue*. The group structure SU(N) gives rise to $N^2 - 1$ generators and therefore 8 strong force carrying gluons (G^a, $a = [1, 8]$) which

	fermions			bosons	
quarks	u	c	t	g	h
	d	s	b	γ	
leptons	ν_e	ν_μ	ν_τ	Z	
	e	μ	τ	W	

Table A.2: Tabular representation particle content of the SM of particle physics. The quarks and leptons are arranged by isospin and generation. The vector gauge bosons and the Higgs scalar boson are located in the right most columns.

interact with coloured objects, the quarks and other gluons. $\mathrm{SU(2)_L}$ mediates the weak isospin, which is exclusive to only the left handed fermions hence the subscript L, here we get 3 vector bosons (W^i, $i = [1, 3]$). Lastly we have $\mathrm{U(1)_Y}$ which only couples to hypercharge, Y with one force carrying boson (B). These gauge bosons are not the typical force carrying bosons that are generally discussed or reported in Tab. A.2. To give rise to these familiar bosons we turn to Higgs mechanism and the spontaneous symmetry breaking (SSB) of the SM gauge symmetry;

$$\mathrm{SU(3)}_c \times \mathrm{SU(2)_L} \times \mathrm{U(1)_Y} \rightarrow \mathrm{SU(3)}_c \times \mathrm{U(1)_{EM}} \, .$$

In the case of SSB in the SM this process is known as electroweak symmetry breaking (EWSB) where the Higgs mechanism (see Section A.3) provides an explanation of the unification of the electromagnetic and weak forces into the electroweak sector at very high energy. At low energy we see the weak and electromagnetic forces as distinct, EWSB gives rise to $\mathrm{U(1)_{EM}}$ which is the group responsible for quantum electrodynamics (QED) and its force carrier the photon (A). EWSB produces the massive $W^\pm$ and Z bosons mediating the weak force via the Higgs mechanism. The SM Lagrangian which encapsulates all the interactions, coupling constants and evolution of the SM fundamental particles can be decomposed into distinct sectors;

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{gauge}} + \mathcal{L}_{\mathrm{fermion}} + \mathcal{L}_{\mathrm{Higgs}} + \mathcal{L}_{\mathrm{Yukawa}} + \mathcal{L}_{\mathrm{CP\,violating}} \, . \tag{A.0.1}$$

A.1 The Gauge Sector

The gauge sector is responsible for the dynamics of the gauge fields and it describes their self interactions,

$$\begin{aligned}\mathcal{L}_{\text{gauge}} &= -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} \\ &= -\frac{1}{4}G^{a,\mu\nu}G^{a}_{\mu\nu} - \frac{1}{4}W^{i,\mu\nu}W^{i}_{\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu}\ ,\end{aligned} \tag{A.1.1}$$

where $G^{a,\mu\nu}$, $W^{i,\mu\nu}$, $B^{\mu\nu}$ are the field strength tensors corresponding to each gauge group. In general for spin one fields we can define the field tensors as,

$$G^{x}_{\mu\nu} = \partial_\mu G^{x}_{\nu} - \partial_\nu G^{x}_{\mu} - g f^{xyz} G^{y}_{\mu} G^{z}_{\nu}\ . \tag{A.1.2}$$

G^{x}_{ν} is a 4-vector bosonic field for each x corresponding to one of the N^2-1 generators, f^{xyz} describes the structure constants which obey following the Lie algebra,

$$[t^a, t^b] = i f^{abc} t^c\ , \tag{A.1.3}$$

where t are the generators of the group and we define g a coupling constant. In SU(3) we get a set of 9 non-zero structure constants f^{abc}:

$$\begin{aligned}&f^{123} = 1\ , \\ &f^{147} = f^{246} = f^{257} = f^{345} = -f^{156} = -f^{367} = \frac{1}{2}\ , \\ &f^{458} = f^{678} = \frac{\sqrt{3}}{2}\ .\end{aligned} \tag{A.1.4}$$

SU(2) the structure constants can be shown to be the fully anti-symmetric tensor, ϵ^{abc}. Lastly, in an Abelian group such as U(1) we know that the generators commute and therefore the structure constants vanish, so we have the following field tenors:

$$\begin{aligned}G^{a}_{\mu\nu} &= \partial_\mu G^{a}_{\nu} - \partial_\nu G^{a}_{\mu} - g_s f^{abc} G^{b}_{\mu} G^{c}_{\nu}\ , \\ W^{a}_{\mu\nu} &= \partial_\mu W^{i}_{\nu} - \partial_\nu W^{i}_{\mu} - g \epsilon^{abc} W^{b}_{\mu} W^{c}_{\nu}\ , \\ B^{a}_{\mu\nu} &= \partial_\mu B_{\nu} - \partial_\nu B_{\mu}\ .\end{aligned} \tag{A.1.5}$$

From Eq. (A.1.1) and Eq. (A.1.5) we can infer insight into the gauge bosons interactions. Firstly, we see that the fields responsible for gluons and weak iso-spin, have

quadratic terms in their fields (e.g. $G^b_\mu G^c_\nu$ and $W^b_\mu W^c_\nu$) which provides a mechanism for self interaction whereas the final field for hypercharge does not. By performing infinitesimal gauge transformations, namely;

$$G^x_\mu \to G^x_\mu - \frac{1}{g}\partial_\mu \alpha^x + f^{xyz}\alpha^y G^z_\mu \,, \tag{A.1.6}$$

we can consider if additional terms in the Lagrangian retain local symmetries in the SM. In particular if we consider adding mass terms to our Lagrangian such as $m^2_G G_\mu G^\mu$ we show that the Lagrangian is no longer gauge invariant, therefore the massive bosons we observe must be introduced via the Higgs mechanism in EWSB.

A.2 The Fermion Sector

The fermionic sector describes the interaction of the quarks and leptons and their interaction with the gauge fields,

$$\mathcal{L}_{\text{fermion}} = i\bar{\mathbf{\Psi}} \not{D} \mathbf{\Psi} \,. \tag{A.2.1}$$

The fermions are represented as Dirac spinors, $\mathbf{\Psi}$ and the dynamics of their interactions are encapsulated in the covariant derivative D_μ,[1]

$$D_\mu = \partial_\mu - ig_Y \frac{Y}{2} B_\mu - ig\frac{\sigma^i}{2} W^i_\mu - ig_s \frac{\lambda^a}{2} G^A_\mu \,. \tag{A.2.2}$$

Here g_Y, g and g_s are the coupling strengths to each of the gauge fields from $\mathrm{U(1)_Y}$, $\mathrm{SU(2)_L}$ and $\mathrm{SU(3)}_c$ respectively. We also introduce the Pauli matrices σ as the generators for $\mathrm{SU(2)_L}$ and the Gell-Mann matrices λ as the generators from $\mathrm{SU(3)}_c$ and note that the generator for $\mathrm{U(1)_Y}$ is the identity thus the fields transform with a scalar multiplier we call hypercharge, Y. Each of the fermion fields has a unique representation under each gauge group, singlet **1**, doublet **2** or triplet **3** (see Tab. A.1). Fermions are divided into quarks and leptons and further by their generation and lastly by left or right handedness. For example an individual u quark (left or right

[1] We note here the usual "slash" notation, $\not{D} = \gamma^\mu D_\mu$ where γ are the gamma Dirac matrices.

Field	$\mathrm{SU(3)}_c$	$\mathrm{SU(2)}_\mathrm{L}$	$\mathrm{U(1)}_\mathrm{Y}$
q_L	**3**	**2**	$1/3$
u_R	**3**	**1**	$4/3$
d_R	**3**	**1**	$-2/3$
l_L	**1**	**2**	-1
ℓ_R	**1**	**1**	-2
ν_R	–	–	–

Table A.3: Fermion field multuplets and their representations or hypercharge for each gauge group.

handed), $u_{\mathrm{L/R}}$ transforms as a triplet under $\mathrm{SU(3)}_c$,

$$u = \begin{pmatrix} u^\mathrm{r} \\ u^\mathrm{g} \\ u^b \end{pmatrix} .$$

The constants r, g, b represent the QCD colour charges. The left handed quarks of a single generation transform as a $\mathrm{SU(2)}_\mathrm{L}$ doublet with an up and down type quark as,

$$q_\mathrm{L} = \begin{pmatrix} u_\mathrm{L} \\ d_\mathrm{L} \end{pmatrix} .$$

Similarly, the left handed leptons of a single generation transform as,

$$l_\mathrm{L} = \begin{pmatrix} \nu_\mathrm{L} \\ \ell_\mathrm{L} \end{pmatrix} .$$

Where we have the charged lepton, ℓ_L and its generational counterpart neutrino, ν_L.

The right handed particles in the SM, u_R, d_R and e_R all transform trivially under $\mathrm{SU(2)}_\mathrm{L}$ and exist as singlets. As alluded to by the omission of entries in Tab. A.3 we in fact do not have a right handed neutrino ν_R in the SM. We can now explicitly write the Lagrangian for the fermion sector for the first generation,

$$\begin{aligned} \mathcal{L}_\mathrm{fermion} = i\bar{q}_\mathrm{L} \not{D} q_\mathrm{L} + i\bar{u}_\mathrm{R} \not{D} u_\mathrm{R} + i\bar{d}_\mathrm{R} \not{D} d_\mathrm{R} + \\ i\bar{l}_\mathrm{L} \not{D} l_\mathrm{L} + i\bar{e}_\mathrm{R} \not{D} e_\mathrm{R} \, . \end{aligned} \tag{A.2.3}$$

The asymmetry between left and right handed particles in the SM has consequences

for mass generation of fermion fields. Writing down the mass term $m\bar{\Psi}\Psi$, we can decompose the field into left and right handed components using the left and right handed projection operators:

$$\Psi_{\mathrm{L}} = P_{\mathrm{L}}\Psi = \frac{1-\gamma^5}{2}\Psi \ , \qquad \Psi_{\mathrm{R}} = P_{\mathrm{R}}\Psi = \frac{1+\gamma^5}{2}\Psi \ , \tag{A.2.4}$$

where $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$. Evaluating the mass term with the left and right decomposition,

$$m\bar{\Psi}\Psi = m(\bar{\Psi}_{\mathrm{L}}\Psi_{\mathrm{R}} + \bar{\Psi}_{\mathrm{R}}\Psi_{\mathrm{L}}) \ . \tag{A.2.5}$$

This demonstrates a mixing of left and right handed fields is required to generate mass via an explicit mass term, this term is in fact not gauge invariant because the left and right fields transform differently. Therefore the mass of our fermions can not be generated in this way and again, we must look to the Higgs Mechanism.

A.3 The Higgs Sector and Higgs Mechanism

The Higgs boson was introduced to the SM to provide a means for many of the SM particles to gain mass in a gauge invariant way. The terms constituting the Higgs sector Lagrangian are the following,

$$\begin{aligned} \mathcal{L}_{\mathrm{Higgs}} &= (D_\mu H)^\dagger (D^\mu H) - V(H) \\ &= (D_\mu H)^\dagger (D^\mu H) - \mu^2 H^\dagger H - \lambda (H^\dagger H)^2 \ , \end{aligned} \tag{A.3.1}$$

where we have the covariant derivative,

$$D_\mu = \partial_\mu - i g_Y \frac{Y}{2} B_\mu - i g \frac{\sigma^i}{2} W^i_\mu \ . \tag{A.3.2}$$

The Higgs field has a scalar potential that depends on two parameters, μ and λ such that if $\mu^2 < 0$ then there exists a minima in the potential at a non-zero value of the field.

Field	$SU(3)_c$	$SU(2)_L$	$U(1)_Y$
H	**1**	**2**	1

Table A.4: Higgs field multiplet and its representations or hypercharge for each gauge group.

To understand the Higgs properties, we first define it as an $SU(2)_L$ doublet made up of two scalar complex fields, ϕ^+ and ϕ^0 in terms of four real scalar fields, ϕ^1, ϕ^2, ϕ^3 and ϕ^4;

$$H = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix} . \tag{A.3.3}$$

Using Eq. (A.3.1) we can show that the potential is minimised for $\mu^2 < 0$ when,

$$|H|^2 = H^\dagger H = -\frac{\mu^2}{2\lambda} = \frac{\nu^2}{2} . \tag{A.3.4}$$

Where we have defined a new quantity ν, the vacuum expectation value (VEV) of the Higgs boson potential and it is experimentally determined to be $\nu \approx 246\,\text{GeV}$ [49]. The Higgs field can be re-written under the unitary gauge where ϕ^3 is chosen to align in the radial direction of the potential and thus h are perturbations uphill from the VEV;

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h \end{pmatrix} . \tag{A.3.5}$$

There is actually flexibility in the gauge choice, the four real fields imply there are in fact four degrees of freedom, three of these correspond to modes which move along the equipotential of the Higgs potential which give rise to three massless Goldstone boson modes. In gauging away these modes the missing degrees of freedom act as the longitudinal modes of the SM gauge bosons giving them mass. The final mode (moving in the radial direction) feels a quadratic potential where perturbations, h are understood to be excitations corresponding to a massive particle, the Higgs boson. We can show by substitution into Eq. (A.3.1) that coefficients in the quadratic terms

of the Higgs boson, h^2 equate to the Higgs boson mass,

$$m_h = \sqrt{2\lambda\nu^2} \ . \tag{A.3.6}$$

After EWSB we can take this gauge choice and expand the covariant derivative term of the Higgs Lagrangian around the Higgs VEV,

$$\begin{aligned}
|D^\mu H|^2 &= \left| \frac{1}{\sqrt{2}} \begin{pmatrix} \partial_\mu - \frac{i}{2}(g_Y B_\mu + g W^3_\mu) & -\frac{ig}{2}(W^1_\mu - iW^2_\mu) \\ -\frac{ig}{2}(W^1_\mu + iW^2_\mu) & \partial_\mu + \frac{i}{2}(-g_Y B_\mu + g W^3_\mu) \end{pmatrix} \cdot \begin{pmatrix} 0 \\ (h+\nu) \end{pmatrix} \right|^2 \\
&= \frac{1}{2} \left| \begin{pmatrix} -\frac{ig}{2}(W^1_\mu - iW^2_\mu)(h+\nu) \\ \partial_\mu(h) + \frac{i}{2}(-g_Y B_\mu + g W^3_\mu)(h+\nu) \end{pmatrix} \right|^2 \\
&= \frac{(\partial_\mu h)^2}{2} + \frac{g^2}{8}(W^1_\mu - iW^2_\mu)(W^{1,\mu} + iW^{2,\mu})(h+\nu)^2 - \\
&\qquad \frac{1}{8}(gW^3_\mu - g_Y B_\mu)(gW^{3,\mu} - g_Y B^\mu)(h+\nu)^2 \ .
\end{aligned} \tag{A.3.7}$$

Eq. (A.3.7) has been written in a deliberately leading way, it is clear to see that there are linear combinations of the massless gauge fields which now exist as quadratic mass terms in the Lagrangian namely,

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp iW^2_\mu) \ , \qquad Z_\mu = \frac{gW^{3,\mu} - g_Y B^\mu}{\sqrt{g^2 + g_Y^2}} = c_W W^{3,\mu} - s_W B^\mu \ . \tag{A.3.8}$$

Where W^+_μ, W^-_μ and Z_μ[1] are our familiar massive force carrying vector bosons. Therefore we can substitute these new fields into Eq. (A.3.7),

$$|D^\mu H|^2 = \frac{(\partial_\mu h)^2}{2} + \frac{g^2}{4} W^-_\mu W^{+,\mu}(h+\nu)^2 - \frac{(g^2 + g_Y^2)}{8} Z_\mu Z^\mu (h+\nu)^2 \ . \tag{A.3.9}$$

The mass terms can now be directly read from Eq. (A.3.9),

$$m_W = \frac{\nu g}{2} \ , \qquad m_Z = \frac{\nu\sqrt{g^2 + g_Y^2}}{2} \ . \tag{A.3.10}$$

It should also be noted there is another orthogonal boson which has no coupling to the Higgs, the photon,

$$A_\mu = c_W W^{3,\mu} + s_W B^\mu \ . \tag{A.3.11}$$

[1]Where $c_W = \cos\theta_W$, $s_W = \sin\theta_W$ and θ_W is the Weinburg weak mixing angle.

As the photon has no coupling to the Higgs we see that it is massless, $m_A = 0$. In summary, the Higgs boson has taken a non-zero VEV which leads to EWSB, the gauge bosons linked to the three broken generators gain mass and the remaining generator associated with $\mathrm{U}(1)_{\mathrm{EM}}$ is unbroken so that the gauge boson retains 0 mass, we now have the more familiar vector gauge bosons mentioned in Tab. A.2 and a complete picture of the SM particles. All that now remains is to explain mass generation for the fermions.

A.4 The Yukawa Sector

As seen in the previous sections, the terms in the Lagrangian so far reviewed that involve fermions do not have a gauge invariant way of generating mass and so again we must turn to the Higgs. We introduce an interaction term between the Higgs boson and the fermions,

$$\mathcal{L}_{\mathrm{Yukawa}} = -(y_u)_{ij}\bar{q}^i_{\mathrm{L}}\tilde{H}u^j_{\mathrm{R}} - (y_d)_{ij}\bar{q}^i_{\mathrm{L}}Hd^j_{\mathrm{R}} - (y_e)_{ij}\bar{l}^i_{\mathrm{L}}He^j_{\mathrm{R}} + \mathrm{h.c.}\,, \tag{A.4.1}$$

where the indices i and j run over the generations of particles. In Eq. (A.4.1) we introduce the conjugate Higgs doublet, $\tilde{H} = i\sigma^2 H^*$ to extract the up type quark fields and we use H as defined in Section A.3 to extract the down type quark fields and leptons. Looking explicitly at the first generation up quark and by substituting in the Higgs doublet and its conjugate,

$$\begin{aligned}\mathcal{L}_{\mathrm{Yukawa}} &= -(y_u)_{11}\bar{q}^2_{\mathrm{L}}\tilde{H}u_{\mathrm{R}} - (y_u)_{11}\bar{u}_{\mathrm{R}}\tilde{H}^\dagger q^2_{\mathrm{L}} \\ &= -\frac{(y_u)_{11}(h+\nu)}{\sqrt{2}}(\bar{u}_{\mathrm{L}}u_{\mathrm{R}} + \bar{u}_{\mathrm{R}}u_{\mathrm{L}}) \\ &= -\frac{(y_u)_{11}h\bar{u}u}{\sqrt{2}} - \frac{(y_u)_{11}\nu\bar{u}u}{\sqrt{2}}\,,\end{aligned} \tag{A.4.2}$$

it is possible to read off the mass term, $m_u = y_u\nu/\sqrt{2}$, where y_u is the Yukawa coupling for the up quark. It is straightforward to repeat this process for the remaining fermions and retrieve similar expressions. The Yukawa matrices y_{ij} are not necessarily diagonal, the flavour (for the weak interactions) and mass basis (for

interactions with the Higgs) of the quarks maybe rotated with respect to one another. Let us introduce two unitary matrices, U and D such that we can define a new object, M;

$$\begin{aligned}(M_u)_{ij} &= U_{\mathrm{L}}^{\dagger}(y_u)_{ij}U_{\mathrm{R}} \\ (M_d)_{ij} &= D_{\mathrm{L}}^{\dagger}(y_d)_{ij}D_{\mathrm{R}}\ .\end{aligned} \tag{A.4.3}$$

These matrices transform the fermion yields from a flavour basis to a mass basis or vice versa. In interactions with neutral currents (those mediated by the photon or Z boson), transformations into a new basis will cancel out due to terms such as $U_{\mathrm{L}}^{\dagger}U_{\mathrm{L}}$ which by the unitary definition equate to the identity. In charged current interactions (those mediated by the W bosons), we get mixing between the up and down type quarks with terms such as $U_{\mathrm{L}}^{\dagger}D_{\mathrm{L}}$. These terms define the Cabibbo-Kobayashi-Mashawa (CKM) matrix which is parameterised by 4 parameters, three angles and a phase [49];

$$V_{\mathrm{CKM}} = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix}\ . \tag{A.4.4}$$

Where $c_{ij} = \cos\theta_{ij}$ and $s_{ij} = \sin\theta_{ij}$ are the sine and cosine of the CKM matrix mixing angles θ_{12}, θ_{13} and θ_{23} and δ the CP violating phase. This parametrisation is accomplished by constraining the CKM matrix by the unitarity property and applying global phases to the quark fields reducing 9 possible parameters in a 3×3 matrix to 4.

A.5 Complete Picture of the Standard Model

For completeness there is one final term in the SM, an additional CP-violating term proportional to the final yet un-discussed free-parameter in the SM.

$$\mathcal{L}_{\text{CP violating}} = \theta_{\mathrm{QCD}}\frac{g_s^2}{32\pi^2}G^a_{\mu\nu}\epsilon_{\mu\nu\sigma\rho}G^{a,\sigma\rho}\ . \tag{A.5.1}$$

This term is gauge invariant therefore permitted in the SM, however current measurements constrain it tightly as being unnaturally small and non-zero, current bounds set it as $|\theta_{\text{QCD}}| \leq 10^{-10}$ [149, 150] and therefore can be assumed to be zero exactly.

Now we have the complete picture of the SM, its particles and their interactions,

$$\begin{aligned}\mathcal{L}_{\text{SM}} = &-\frac{1}{4}F^{\mu\nu}F_{\mu\nu} \\ &+ i\bar{\mathbf{\Psi}}\not{D}\mathbf{\Psi} + \text{h.c.} \\ &+ \bar{\mathbf{\Psi}}_i y_{ij}\mathbf{\Psi}_j\phi + \text{h.c.} \\ &+ |(D_\mu\phi)|^2 - V(\phi)\ .\end{aligned} \tag{A.5.2}$$

A.6 Beyond the Standard Model

The SM is a very complete picture explaining many of the fundamental particles and their interactions to unprecedented precision. However, as briefly outlined at the start of this appendix there exist a number of short comings in the SM.

- Neutrino oscillations: SM neutrinos are only left handed which therefore means they do not have mass in the SM, right handed components are necessary for mass generation from the Yukawa sector. However experimental observations of neutrino oscillations imply that they do in fact have non-zero masses [151–153].

- Cosmological measurements: Velocities of galaxies in clusters and rotational curves of stars contained within galaxies imply greater mass exists than is visible and explainable by the 3 fundamental forces of nature of the SM [154, 155]. This undetectable mass is coined "Dark Matter" and it is calculated to take up roughy $\approx 26\%$ of the energy content of the universe. Further to this measurements of the rate of the expansion of the universe requires an introduction of a new form of energy, Dark Energy [156, 157] which is $\approx 68\%$ of the energy content of the universe.

- The CP problem: The vast majority of the observable matter content of the universe is matter, not antimatter. There exist CP violating processes which

treat matter and antimatter differently within the SM however, these processes are not enough to explain the disparity between the amount of matter and antimatter that exists in the present day [158].

- Flavour problem: There is no natural explanation of the size and scaling of CKM parameters or Yukawa couplings between each generation of matter within the SM [159]. These parameters are only defined by experimental observation.

- Gravity: There exists no force carrying boson for gravity within the SM.

This summary of a selection of problems is not to discredit the SM, but it is clear there is room for improvement and additions to the SM of physics. In this thesis we will discuss methods to more tightly constrain the Yukawa coupling to the charm quark and such methods might be applied to probe other parameter constraints in the SM, so that we might find hints to BSM physics or discover new tensions between theory and reality.

Appendix B

Anomaly Detection at Hartlepool Power Station Dashboard

During the placement at Hartlepool Power Station a GUI dashboard was developed in which engineers at the plant can test and understand the tool chains developed. The information in this appendix is the documentation provided which details the dashboards use.

The dashboard provided is intended for use of an engineer to give an idea of changes in the operational parameters. To that end there needs to be some manual interpretation and set up. The dashboard has been developed with the condenser and vacuum pump systems in mind however can be extended to any other set of systems. To ensure correct functionality the following points need to be kept in mind:

1. We require at around $\sim$100 data points per window to run α analysis. e.g. A week for hourly sampling,

2. Systems may be correlated,

3. Systems which do not feature strong periodicity will not benefit from α analysis,

4. The rates thresholds will need refitting on any unseen dataset, but the dashboard will still function omitting the threshold line i.e. Using $\bar{R}_F(\alpha_{\text{crit}})$ instead

of $\bar{R}_F(\alpha_{\text{crit}}, \bar{\beta}_{\text{crit}})$.

1. The α analysis outlined in this report requires an initial week of data to fit the first $\alpha(t)$ with hourly sampling. The Fourier transform and χ^2 fit requires enough data points to reasonably understand the underlying periodicity in the data and get a good fit on $\alpha(t_0)$.

2. Systems may be correlated, one such example is the CGO data. The thermocouple temperature "RP209" in the channels of the reactor depend most closely on two other features: the power output of the reactor "BF603" and the gag control depth "GA201". If we naively plug in the thermocouple data alone we neglect to consider they are coupled to a changes in the gag or power output. Therefore sensible combinations of variables should be used. Running the thermocouple, gag and power through the PCA model together gives us a better understanding of which features are changing independently of other features by using the marginalised T^2 and Q statistic.

3. If a feature has low periodicity then there will be no significant benefit in running the α analysis. This is particularly true of the CGO dataset in which there are long linear trend changes but no daily, weekly or yearly behaviour. This is why the dashboard provides a "raw" option which skips the α analysis step entirely and only takes first order differences.

4. The threshold line is optimized for each family of features e.g. "CC". If an optimal fit does not exist in the parameters file *Vars.py* then the mean rolling threshold line will be omitted. One can be added by adding an entry into the parameters file.

The dashboard is designed to make the analysis as customisable, intuitive and easy to navigate as possible. To this end, the graphs are all fully interactive and each tab contains a read out of key information of the plots shown. The dashboard also creates a cache of processed data and graphs to prevent needlessly recomputing the same data. To load data into the dashboard simply drop .xlsx files into "data/RawData". The .xlsx files must be named for the variable e.g. "R1 BF603.xlsx" and there must

be three columns, in this case “Timestamps”, “R1 BF603” and “Error” which contain the PI explorer timestamp, the values of R1 BF603 and the standard deviations of R1 BF603 respectively. The dashboard will comes with demonstration data with the correct format as an example. This tool was constructed using Dash and Plotly in Python3, the installation of required packages can be performed with pip, the default package installer for PYTHON. There will be provided install script which will function on Windows and Linux machines.

The dashboard is sub-divided into a set of tabs namely:

1. Variables,
2. Fourier Analysis,
3. Alpha Analysis,
4. BOCPD,
5. PCA.

Tabs 1 and 2 are intended for data exploration and visualization. *Variables* provides plots of the raw data provided in the “RawData” folder and a basic read back. *Fourier Analysis* provides some insight into the mathematics underling in the α analysis. The remaining tabs 3, 4 and 5 run some form of analysis and have some dependencies on each other e.g. BOCPD analysis will be unavailable for α data unless the *Alpha Analysis* tab has run the required processing for data you wish to process. The dashboard is also able to keep a log of graphs or runs details that are performed to keep a record of when certain combinations of features where checked or produced. These are saved in the “cachefiles/” folder.

The rest of this section will briefly explain the operation and features of each tab and any dependencies.

B.1 Variables

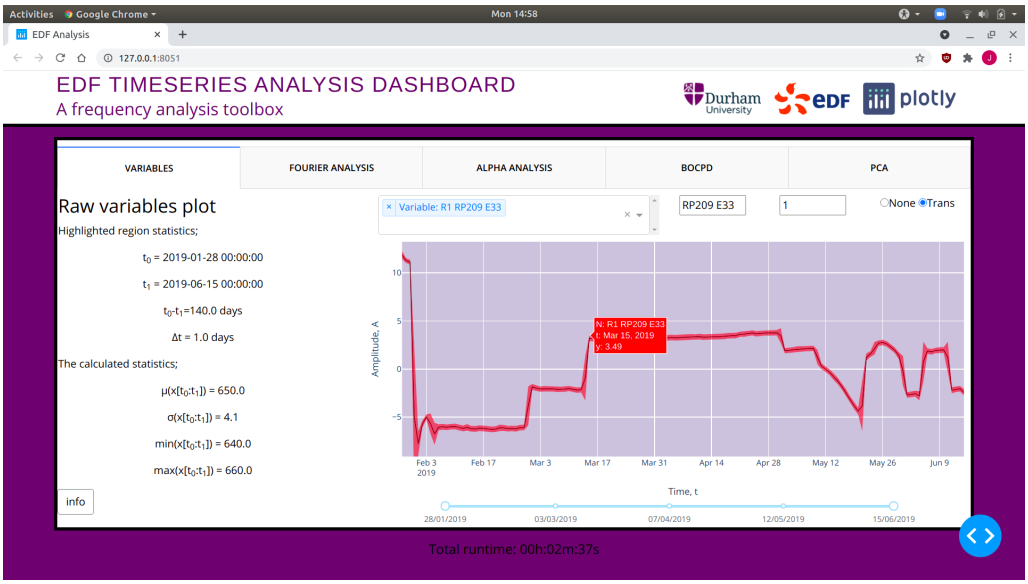


Figure B.1: Screenshot of tab *Variables*

The *Variables* tab provides an overview of the raw features being fed into the dashboard for quick sanity check on which features are loaded in. On the left hand side there is a readout summarizing the highlighted regions start time t_0, end time t_1, the duration and time step between readings. The dashboard automatically detects units in seconds, minutes, hours and days. It also calculates the mean μ, standard deviation σ, minimum and maximum values. The selector labelled "None" and "Trans" mean centres the features for easier comparison between features which have a large difference their means.

Features on this tab include:

- dropdown feature selector,
- input feature string search,
- input errors factor multiplier,
- selector to re-scale features,

- interactive graph, hover mouse for info and zoom. Double click to reset,
- slider to highlight region of graph in purple,
- on left info of key statistics over the region highlighted,
- info button explaining tab usage,

and dependencies:

- .xlsx files in the "data/RawData" folder.

B.2 Fourier Analysis

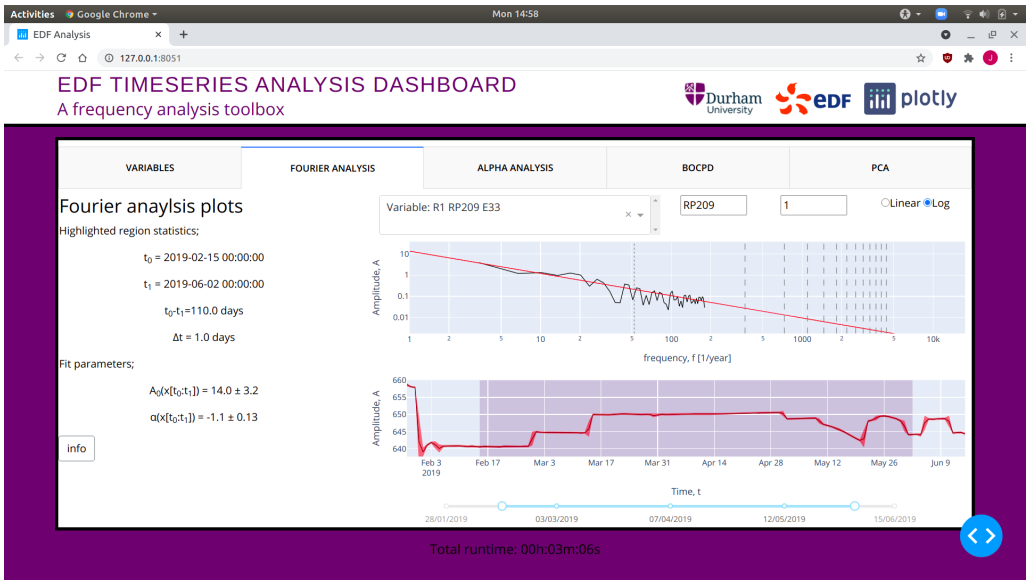


Figure B.2: Screenshot of tab *Fourier Analysis*

The *Fourier Analysis* tab provides an demonstration the α analysis fit. It is possible to adjust with the window size with is indicated by the highlighted region and the minimum frequency of the fit. These changes will re-perform the calculation and display the Fourier transform in the top graph and the power law spectrum red with a red line. On the left hand side there is a readout summarizing the fit on the

highlighted region and general information. Lastly, the Fourier spectrum graph has vertical lines corresponding to key frequencies we might expect which is harmonics in weekly or daily oscillations.

Features on this tab include:

- dropdown feature selector,
- input feature string search,
- input errors factor multiplier,
- linear or log y-axis scale selector,
- interactive graphs, hover mouse for info and zoom. Double click to reset,
- slider to highlight region of graph in purple. Region used for fitting frequency spectrum,
- on left info of key statistics over the region highlighted,
- info button explaining tab usage,

and its dependencies;

- .xlsx files in the "data/RawData" folder.

B.3 Alpha Analysis

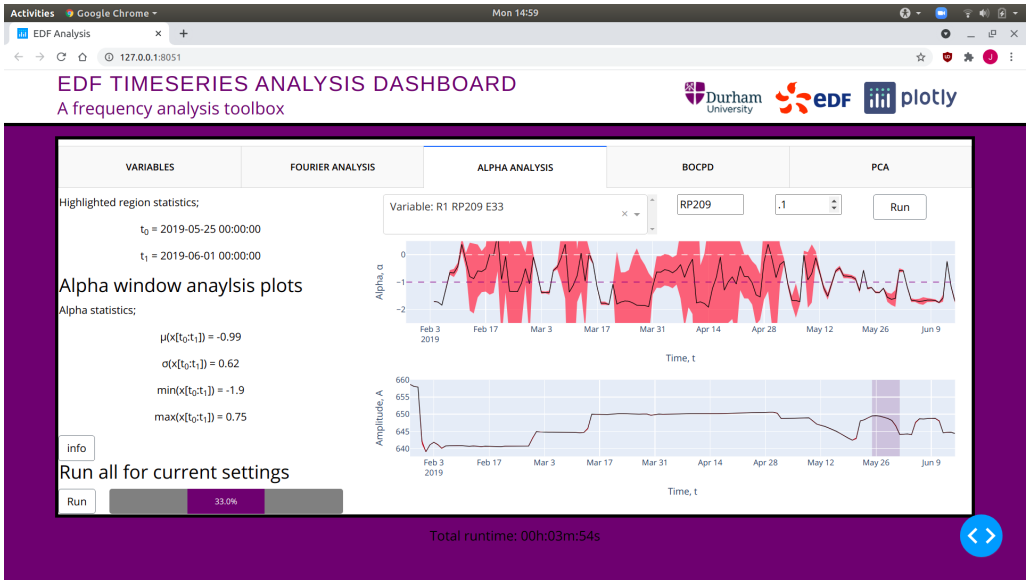


Figure B.3: Screenshot of tab *Alpha Analysis*

The *Alpha Analysis* tab has the functionality to run the α analysis over the chosen feature or on all features by clicking the Run button. The run button loops through all the files containted within the "data/RawData" folder and produces the corresponding α .xlsx sheets. To process one feature select the feature to analyse and click Run at the top. Takes roughly around one minute per event depending on the size of the .xlsx file. A load icon and the progress bar will appear update to show the user that code is running behind the scenes. The dashboard will save results in "data/AlphaData". The size of fitting the window can be adjusted in the *Var.py* file the default is 7 days. Some key statistics are calculated and displayed on the left about the α series produced. Lastly clicking on a α data point in the upper graph will highlight a region in the lower graph showing which window produced the α.

Features on this tab include:

- dropdown feature selector,
- input feature string search,

- input errors factor multiplier for graphs,
- run button on selected feature,
- interactive graphs, hover mouse for info and zoom. Double click to reset,
- alpha graph data points clickable to highlight region of raw feature that data point is calculated from,
- on left info of key statistics over the region highlighted,
- info button explaining tab usage,
- run button. Loops over all Dropdown options and produces .xlsx files in "data/AlphaData". Progress bar updates every 5 seconds,

and its dependencies:

- .xlsx files in the "data/RawData" folder.

B.4 BOCPD

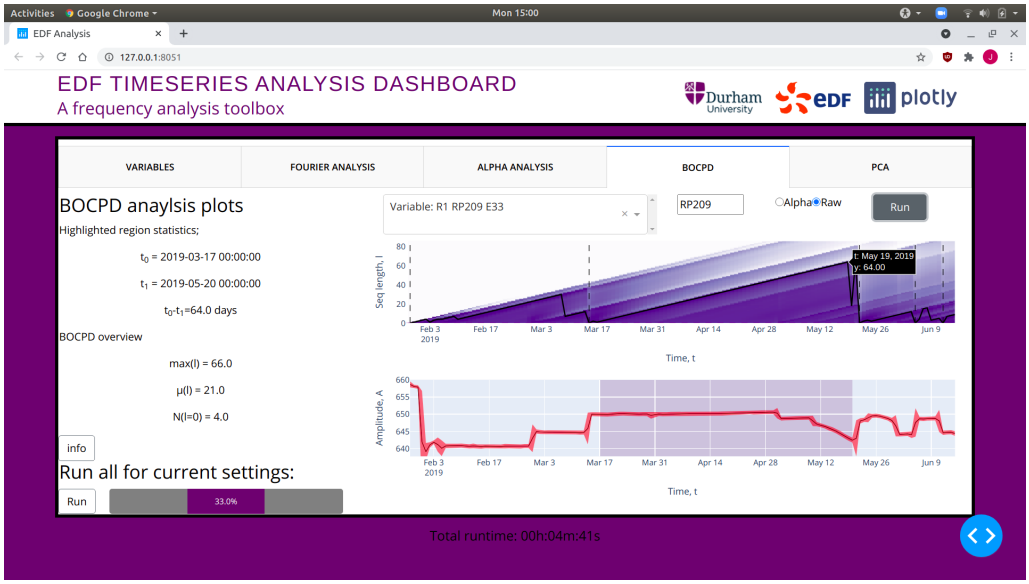


Figure B.4: Screenshot of tab *BOCPD*

The upper graph shows the BOCPD heat map and the lower plot the feature fed into the BOCPD analysis. In the top plot the heatmap represents the $\boldsymbol{L}(r,t)$ probability matrix where the darker highlighted regions indicate points on the plot which are the most likely and the white regions are vanishingly unlikely. Clicking on this plot will highlight what sequence in the feature produces this run length. The thick black line is the most probable sequence length at this particular time where this line drops to 0 we plot a vertical line. This is a strongly defined CP. Much like the *Alpha Analysis* tab this tab has the functionality to run through all the "data/RawData" or "data/AlphaData/" folders' .xlsx files and produce the corresponding BOCPD .xlsx sheets. Simply select the feature and click Run which takes around 1 minute per event depending on the length of the region. A load icon and the progress bar will appear update to show the user that code is running behind the scenes. Some key statistics are calculated and displayed on the left. Lastly, clicking on a sequence length data point in the upper graph will highlight a region in the lower graph showing which data would consist on a sequence of that length.

Features on this tab include:

- dropdown feature selector,
- input feature string search,
- "Alpha" or "Raw" selection to run data from "data/RawData" or "data/AlphaData",
- run button on selected feature,
- interactive graphs, hover mouse for info and zoom. Double click to reset,
- BOCPD graph data points clickable to highlight region of sequence for which a series is that long,
- on left info of key statistics over the region highlighted,
- info button explaining tab usage,

- run button. Loops over all dropdown options and produces .xlsx files in "data/RawData" or "data/AlphaData". Progress bar updates every 5 seconds,

and its dependencies:

- .xlsx files in the "data/RawData" folder for "Raw" option,
- .xlsx files in the "data/AlphaData" folder for "Alpha" option.

B.5 PCA

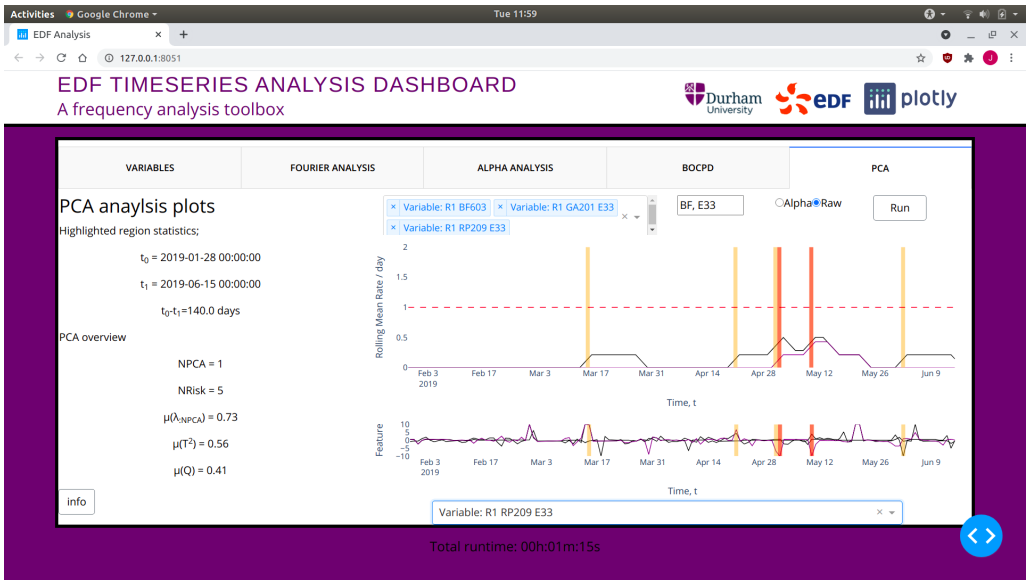


Figure B.5: Screenshot of tab *PCA*

The *PCA* tab takes either "Raw" or "Alpha" data from "data/RawData" or "data/AlphaData" folders which can be selected using the multiselect dropdown menu. Click Run to perform the analysis, this will take around 30/40 seconds to complete depending on the length of the data. We then produce two plots, the top shows the rolling rates. The lower plot shows the normalised distance in standard deviations of the PCA fitted features from their mean. A distance of two or more from zero is unlikely in the model and five or more is exceedingly unlikely. Any regions flagged as

anomalous in both T^2 and Q will appear in yellow or red. Red for the marginalised feature selected in the bottom dropdown menu is shown to violate the threshold on its own and yellow if the threshold is violated in general across all features. Some key statistics are displayed on the left, the two to monitor most closely are $\mu(\lambda)$ which is the mean fractional variance in the kept PCA components, if this value drops below around 0.6 consider adding extra PCA components through the NPCA variable in *Var.py*. The other to monitor is NRisk which is the number of flagged anomalous regions which counts the data points violating the thresholds.

Features on this tab include:

- dropdown feature selector,
- input feature string search,
- "Alpha" or "Raw" selection to run data from "data/RawData" or "data/AlphaData" folders,
- run button. Runs the PCA rolling window analysis for the selection,
- interactive graphs, hover mouse for info and zoom. Double click to reset,
- on left info of key statistics over the region highlighted,
- info button explaining tab usage,
- dropdown feature selector for marginal feature comparison,

and its dependencies:

- .xlsx files in the "data/RawData" folder for "Raw" option,
- .xlsx files in "data/AlphaData" folder for "Alpha" option.

B.6 Changing Default Parameters

In the dashboard install directory there is a file *Var.py* which can be edited to adjust various parameters involved in analysis or dashboard functionality. It is recommended many parameters are left alone as they have been carefully tuned except when the file provides further information.

Appendix C

Social Mixing Matrices at Cox's Bazar

In this appendix we detail supplementary information from the placement with UN Global Pulse. Appendix C.1 details the released contact matrices and the available open source code used to derive them. Appendix C.2 details the contact survey which was performed within the camp which provides the information used to derive the aggregate interaction matrices implemented in June-Cox. Lastly, Appendix C.3 details the questions posed to the Community Based Protection team and the information they provided which was used to tune the rules and social patterns within June-Cox.

C.1 Code and Availability

- June and June-Uk: The current public release of the June simulation framework and by extension the latest version of the June-Uk model, can be found at `https://github.com/IDAS-Durham/JUNE`
- June-Cox: The current public release of June-Cox epidemic model can be found at `https://github.com/UNGlobalPulse/UNGP-settlement-modelling`

- Data: The data from the survey is available by application at `https://microdata.unhcr.org/index.php/catalog/587`
- Contact Survey: Details and calculation at `https://github.com/UNGlobalPulse/UNGP-contact-survey`
- JUNE-UK Household contact matrix: Details and calculation at `https://github.com/IDAS-Durham/june_household_matrix_calculation`
- Contact Matrix Results: Our contact matrices are reported here `https://github.com/IDAS-Durham/june_mixed_method_CM_results` formatted in excel documents for convenience.

C.2 Survey

The survey between October-November 2020 was conducted by enumerators from the UNHCR Community Based Protection team who regularly conduct surveys within the settlement following standard UNHCR practices [132, 133]. Data was collected from 22 camps in the Kutapalong-Balukhali Expansion Site (part of the Cox's Bazar refugee settlement) consisting of 2 men and 2 women in each of the following categories: < 18 years; ≥ 18 years < 60; ≥ 60 years. In addition 2 persons with disabilities were surveyed to make a total of 308 respondents. Anonymised results and additional metadata, can be accessed through UNHCR [134].

The survey was conducted by enumerators randomly sampling households in each camp and visiting them in person. Only one respondent per household was permitted and responses were collected using the Kobo Toolbox [160] based on the Open Data Kit [161]. The survey was formatted as follows (italicised text is spoken):

This questionnaire has been designed by teams from United Nations Global Pulse and UNHCR and is to inform efforts to better understand how people move around in the camp and interact with others to better understand how COVID-19 might spread in the camp to inform future COVID-19 protection measures.

Good day by name is _____ from UNHCR and I am here to conduct a survey. This study is part of a scientific research project from United Nations Global Pulse and UNHCR. In this study, we will ask questions to better understand how people move around in the camp and interact with others. Your decision to complete this study is completely voluntary and you may decline to answer at any time. Your answers will be completely anonymous. The results of the research may be presented at scientific meetings or published in scientific journals. For any questions or comments please contact: _____. The survey should not take longer than 30 minutes.

1. - **If adult:** *Do you declare that you are at least 18 years of age and that you agree to complete this survey voluntarily?*
   - **If child:**
     - **To parent or guardian:** *Do you declare that you are at least 18 years of age, that you are the parent or guardian of this child and that you give consent for your child to complete this survey voluntarily?*
     - **To child:** *Do you declare that this is your parent or guardian and that you give consent to complete this survey voluntarily?*
2. *Sex:* Female, Male, Other, Do not want to answer
3. *Location at the time:* _____ (camp)
4. *Age:* under 18, over 18 but under 60, over 60
5. *Disability:* Y/N
6. *Do you have access to a face mask?* Y/N
7. *When the learning centres were open, did you attend any formal education?* Y/N
8. - **If yes:**

(a) *When you attended formal education, how much time do you spend there?* 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, other (please specify)

(b) *When you attended formal education, approximately how many children do you come into contact with (for example, talk to)?*

(c) *When you attended formal education, approximately how many adults do you come into contact with (for example, talk to)?*

9. *Do you ever go to a food distribution center?* Y/N

10. • **If yes:**

(a) *When you go to a food distribution center, how much time do you spend there?* 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, other (please specify)

(b) *When you go to a food distribution center, approximately how many children do you come into contact with at the center (for example, talk to)?*

(c) *When you go to a food distribution center, approximately how many adults do you come into contact with at the center (for example, talk to)?*

(d) *When you go to the food distribution center, do you wear a mask in the center?*

11. *Do you ever go to an e-voucher outlet?* Y/N

12. • **If yes:**

(a) *When you go to an e-voucher outlet, how much time do you spend there?* 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, other (please specify)

(b) *When you go to an e-voucher outlet, approximately how many children do you come into contact with at the outlet (for example, talk to)?*

(c) *When you go to an e-voucher outlet, approximately how many adults do you come into contact with at the outlet (for example, talk to)?*

(d) *When you go to an e-voucher outlet, do you wear a mask in the outlet?*

13. *Do you ever go to a community center?* Y/N

14. • **If yes:**

(a) *When you go to a community center, how much time do you spend there?* 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, other (please specify)

(b) *When you go to a community center, approximately how many children do you come into contact with at the center (for example, talk to)?*

(c) *When you go to a community center, approximately how many adults do you come into contact with at the center (for example, talk to)?*

(d) *When you go to a community center, do you wear a mask in the center?*

15. *Do you ever go to a religious meeting?* Y/N

16. • **If yes:**

(a) *When you go to a religious meeting, how much time do you spend there?* 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, other (please specify)

(b) *When you go to a religious meeting, approximately how many children do you come into contact with at the meeting (for example, talk to)?*

(c) *When you go to a religious meeting, approximately how many adults do you come into contact with at the meeting (for example, talk to)?*

(d) *When you go to a religious meeting, do you wear a mask in the meeting?*

17. (a) *When you go to a water pump or latrine, how much time do you spend there?* 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, other (please specify)

    (b) *When you go to a water pump or latrine, approximately how many children do you come into contact with (for example, talk to)?*

    (c) *When you go to a water pump or latrine, approximately how many adults do you come into contact with (for example, talk to)?*

    (d) *When you go to a hand pump or latrine, do you wear a mask?*

C.3 Questions for the CBP team

To supplement our analysis, a series of informal interviews were conducted with members of the Cox's Bazar refugee settlement UNHCR Community Based Protection (CBP) team. In each of these interviews a set of general enquires into the behaviour and attendance rates were asked of members of the protection team which worked closely with those venue types.

This questionnaire has been designed by teams from United Nations Global Pulse and UNHCR and is to inform efforts to better understand how people engage with each venue in the camp and the demography of the venues.

For the following venues: Community centres, Female friendly spaces, Food distribution centres, E-voucher outlets, Non-food distribution centres – including LPG and blanket centres – Religious centres and Learning centres. Where you are able and suitably informed please could you answer the following questions;

1. • *Can you describe what a day looks like at* **venue** *?*
     - *How many people do you expect at minimum and peak times?*
     - *How do these days and numbers of people vary by day, week, month/season?*
     - *Why do you think there are these variations?*

2. • *What is the makeup of multigenerational households – are there generally three generations or more?*
   - *Do these households include extended family?*
   - *Is this a cultural issue or a space constraint?*

3. • *What age do children typically move through the camp independently?*
   - *Move out from parents shelter?*
   - *Go to* **venues** *on their own? (e.g collect items from the distribution centres for their shelter)*
   - *How many hours do they spend moving around in the camp independently?*
   - *Who do they mostly have contact with when they move around? (e.g. more children, teachers at school, other adults? all)*

4. • *What time do* **venues** *close?*

Here we outline the key findings from the interviews used to define the virtual venues in June-Cox.

- Community centres and Female friendly spaces:
  - *Busiest in morning.*
  - *Typically 35-40 people per day.*
  - *Closed Friday and Saturday.*
  - *Less busy during rainy season and religious holidays but more busy in national holidays.*
- Distribution centres (Food and Non-Food):
  - *Typically 300-400 people per day.*
  - *Sunday busiest day 500 people per day.*

  - *Families permitted to collect food every two weeks.*
  - *Children not permitted on their own.*

- E-voucher:
  - *Same as Distribution Centers.*
  - *No limit on frequency of attendance.*

- Learning Centers:
  - *Children (4-13) attend in morning or afternoon slots of 2-3 hours.*
  - *Typically 80-100 children per group.*
  - *Closed Friday and Saturday*
  - *Less busy during rainy season.*

- Religious Centers:
  - *Attended only by men 11+ in age.*
  - *Five daily prayers throughout the day.*
  - *Typically 150-200 people per day.*
  - *Friday busiest day 200-300 people.*
  - *Less busy during rainy season but more busy in religious holidays 600-700.*

Bibliography

[1] J. Walker and F. Krauss, *Constraining the Charm-Yukawa coupling at the Large Hadron Collider*, *Physics Letters B* **832** (2022) .

[2] J. Walker, J. Aylett-Bullock, D. Shi, A. G. K. Maina, E. S. Evers, S. Harlass et al., *A Mixed-Method Approach to Determining Contact Matrices in the Cox's Bazar Refugee Settlement* , *Royal Society Open Science* **10** (2023) .

[3] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.

[4] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*. Addison-Wesley, 1995.

[5] J. Campbell, J. Huston and F. Krauss, *The Black Book of Quantum Chromodynamics : a Primer for the LHC Era*. Oxford University Press, 2018.

[6] R. K. Ellis, W. J. Stirling and B. R. Webber, *QCD and Collider Physics*. Cambridge University Press, 2011.

[7] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Erratum to: Asymptotic formulae for likelihood-based tests of new physics*, *European Physical Journal C* **71** (July, 2010) .

[8] Read A. L. , *Presentation of search results: the CLs technique*, *J. Phys. G Nucl. Part. Phys.* **28** (2002) 2693–2704.

[9] UN Global Pulse, "UN Global Pulse Website." `https://www.unglobalpulse.org/`, 2023.

[10] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Q. Al-Dujaili, Y. Duan, O. Al-Shamma et al., *Review of deep learning: concepts, cnn architectures, challenges, applications, future directions*, *Journal of Big Data* **8** (2021) .

[11] R. A. Fisher, "Iris." UCI Machine Learning Repository, 1988.

[12] V. Gligorov, *Real-time data analysis at the LHC: present and future*, *J. Mach. Learn. Res.* **42** (2015) 1–18.

[13] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel et al., *Machine learning at the energy and intensity frontiers of particle physics* , *Nature* **560** (2018) 41–48.

[14] Schwartz, Matthew D., *Modern Machine Learning and Particle Physics*, *Harvard Data Science Review* **3** (May, 2021) .

[15] He K., Zhang X., Ren S. and Sun J., *Deep Residual Learning for Image Recognition*, `1512.03385`.

[16] M. J. F. T Villegas and M. Rodríguez, *Principal component analysis for fault detection and diagnosis. Experience with a pilot plant* , *International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics - Proceedings* (Jan., 2010) .

[17] D. Garcia-Alvarez, *Fault Detection Using Principal Component Anaylsis (PCA) In A Wastewater Treatment Plant (WWTP)* , .

[18] B. De Ketelaere, M. Hubert and E. Schmitt, *Overview of PCA-Based Statistical Process-Monitoring Methods for Time-Dependent, High-Dimensional Data* , *Journal of Quality Technology* **47** (Oct., 2015) 318–335.

[19] C. Chatfield and A. Collins, *Introduction to Multivariate Analysis*. Feb., 2018.

[20] J. E. Jackson and G. S. Mudholkar, *Control procedures for residuals associated with principal component analysis*, *Technometrics* **21** (1979) 341–349.

[21] G. J. J. van den Burg and C. K. I. Williams, *An Evaluation of Change Point Detection Algorithms*, 2003.06222.

[22] R. P. Adams and D. J. C. MacKay, *Bayesian Online Changepoint Detection*, 0710.3742.

[23] J. Aylett-Bullock, C. Cuesta-Lazaro, A. Quera-Bofarull, M. Icaza-Lizaola, A. Sedgewick, H. Truong et al., *June: open-source individual-based epidemiology simulation*, *Royal Society Open Science* **8** (2021) .

[24] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, *Eur. Phys. J. C* **76** (2016) , [1512.01178].

[25] M. Bähr, S. Gieseke, M. A. Gigg, D. Grellscheid, K. Hamilton, O. Latunde-Dada et al., *Herwig physics and manual*, *The European Physical Journal C* **58** (Nov., 2008) 639–707.

[26] C. Bierlich, S. Chakraborty, N. Desai, L. Gellersen, I. Helenius, P. Ilten et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, *SciPost Phys. Codebases* (2022) .

[27] Gleisberg T. et al., *Event generation with SHERPA 1.1*, *JHEP* **2** (2009) , [0811.4622].

[28] Gleisberg T. and Höche S., *Comix, a new matrix element generator*, *JHEP* **12** (2008) , [0808.3674].

[29] F. Krauss, R. Kuhn and G. Soff, *AMEGIC++ 1.0: A Matrix element generator in C++*, *JHEP* **02** (2002) , [hep-ph/0109036].

[30] NNPDF Collaboration, *Parton distributions for the LHC run II*, *JHEP* **2015** (2015) .

[31] Buckley A. et al., *LHAPDF6: parton density access in the LHC precision era*, *Eur. Phys. J. C* **75** (2015) , [1412.7420].

[32] R. Kleiss, W. Stirling and S. Ellis, *A new monte carlo treatment of multiparticle phase space at high energies*, *Computer Physics Communications* **40** (1986) 359–373.

[33] Schumann S. and Krauss F., *A Parton shower algorithm based on Catani-Seymour dipole factorisation*, *JHEP* **3** (2008) , [0709.1027].

[34] G. Altarelli and G. Parisi, *Asymptotic freedom in parton language*, *Nuclear Physics B* **126** (Aug., 1977) 298–318.

[35] V. N. Gribov and L. N. Lipatov, *Deep inelastic e p scattering in perturbation theory*, *Sov. J. Nucl. Phys.* **15** (1972) 438–450.

[36] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.*, *Sov. Phys. JETP* **46** (1977) 641–653.

[37] B. Webber, *A QCD model for jet fragmentation including soft gluon interference*, *Nuclear Physics B* **238** (1984) 492–528.

[38] T. Sjostrand and M. van Zijl, *A Multiple Interaction Model for the Event Structure in Hadron Collisions*, *Phys. Rev. D* **36** (1987) .

[39] S. Ovyn, X. Rouby and V. Lemaitre, *DELPHES, a framework for fast simulation of a generic collider experiment*, 0903.2225.

[40] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley and D. S. Burke, *Strategies for mitigating an influenza pandemic*, *Nature* **442** (July, 2006) 448–452.

[41] D. L. Chao, M. E. Halloran, V. J. Obenchain and I. M. Longini, Jr, *Flute, a publicly available stochastic influenza epidemic simulation model*, *PLOS Computational Biology* **6** (Jan., 2010) 1–8.

[42] J. Aylett-Bullock, C. Cuesta-Lazaro, A. Quera-Bofarull, A. Katta, K. H. Pham, B. Hoover et al., *Operational response simulation tool for epidemics within refugee and IDP settlements: A scenario-based case study of the Cox's Bazar settlement* , *PLOS Computational Biology* **17** (Oct., 2021) .

[43] P. E. M. Fine and J. A. Clarkson, *Measles in England and Wales—I: An Analysis of Factors Underlying Seasonal Patterns* , *International Journal of Epidemiology* **11** (Mar., 1982) 5–14.

[44] R. M. Anderson and R. M. May, *Age-related changes in the rate of disease transmission: implications for the design of vaccination programmes* , *Epidemiology and Infection* **94** (June, 1985) 365–436.

[45] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk et al., *Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases* , *PLOS Medicine* **5** (Mar., 2008) .

[46] L. Fumanelli, M. Ajelli, P. Manfredi, A. Vespignani and S. Merler, *Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread* , *PLoS Computational Biology* **8** (Sept., 2012) .

[47] S. Y. Del Valle, J. M. Hyman, H. W. Hethcote and S. G. Eubank, *Mixing patterns between age groups in social networks*, *Social Networks* **29** (Oct., 2007) 539–554.

[48] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk et al., *Social contacts and mixing patterns relevant to the spread of infectious diseases*, *PLOS Medicine* **5** (Mar., 2008) .

[49] Particle Data Group collaboration, R. L. Workman and Others, *Review of Particle Physics*, *PTEP* **2022** (2022) .

[50] W. W. Armstrong et al., *ATLAS: Technical proposal for a general-purpose p p experiment at the Large Hadron Collider at CERN*, tech. rep., ATLAS Collaboration, Dec., 1994.

[51] CMS collaboration, *CMS, the Compact Muon Solenoid: Technical proposal*, tech. rep., CMS Collaboration, Dec., 1994.

[52] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC* , *Phys. Lett. B* **716** (2012) 1–29.

[53] CMS Collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC* , *Phys. Lett. B* **716** (2012) 30–61, [1207.7235].

[54] B. Schmidt, *The High-Luminosity upgrade of the LHC: Physics and Technology Challenges for the Accelerator and the Experiments* , *Journal of Physics: Conference Series* **706** (Apr., 2016) .

[55] J. Pequenao and P. Schaffner, "How ATLAS detects particles: diagram of particle paths in the detector." 2013.

[56] ATLAS collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) .

[57] W. Waltenberger and R. Frühwirth, *Adaptive Vertex Fitting*, *Journal of Physics G: Nuclear and Particle Physics* **34** (Nov., 2007) .

[58] M. Padilla, *Measurement of The Single Top Quark Production Cross Section at the Square Root of s =1.96 TeV*, Ph.D. thesis, FNAL, Jan., 2011.

[59] R. Frühwirth, *Application of Kalman filtering to track and vertex fitting*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **262** (1987) 444–450.

[60] S. Catani, Y. Dokshitzer, M. Seymour and B. Webber, *Longitudinally-invariant k-clustering algorithms for hadron-hadron collisions* , *Nuclear Physics B* **406** (1993) 187–224.

[61] CMS collaboration, *A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging*, tech. rep., CERN, Geneva, 2009.

[62] Cacciari M. et al., *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) , [0802.1189].

[63] J. Gallicchio and M. Schwartz, *Seeing in Color: Jet Superstructure*, *Physical review letters* **105** (July, 2010) .

[64] A. Larkoski, J. Thaler and W. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, *Journal of High Energy Physics* **2014** (Aug., 2014) .

[65] S. V. Chekanov and J. Proudfoot, *Searches for TeV-scale particles at the LHC using jet shapes*, *Phys. Rev. D* **81** (June, 2010) .

[66] L. Almeida, S. Lee, G. Perez, G. Sterman, I. Sung and J. Virzi, *Substructure of high-p_T Jets at the LHC*, *Physical Review D* (July, 2008) .

[67] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) , [1011.2268].

[68] LHC Higgs Cross Section Working Group, *LHC HXSWG interim recommendations to explore the coupling structure of a Higgs-like particle* , tech. rep., CERN, 2012.

[69] Evans L. and Bryant P., *LHC machine*, *J. Instrum.* **3** (2008) .

[70] Particle Data Group, *Review of Particle Physics*, *Progress of Theoretical and Experimental Physics* **2020** (2020) .

[71] ATLAS Collaboration, *Measurement of the production cross section for a Higgs boson in association with a vector boson in the $H \to WW* \to l\nu l\nu$ channel in pp collisions at s=13 TeV with the ATLAS detector*, *Phys. Lett. B* **798** (2019) .

[72] CMS Collaboration, *Constraints on anomalous Higgs boson couplings to vector bosons and fermions in its production and decay using the four-lepton final state*, *Phys. Rev. D* **104** (2021) , [2104.12152].

[73] ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector* , *Phys. Lett. B* **784** (2018) 173–191.

[74] CMS Collaboration, *Measurement of the Higgs boson production rate in association with top quarks in final states with electrons,muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **81** (2020) , [2011.03652].

[75] ATLAS Collaboration, *Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **99** (2019) .

[76] CMS collaboration, *Observation of the Higgs boson decay to a pair of tau leptons with the CMS detector* , *Physics Letters B* **779** (2018) 283–316.

[77] ATLAS Collaboration, *Measurement of the associated production of a Higgs boson decaying into b-quarks with a vector boson at high transverse momentum in pp collisions at s=13TeV with the ATLAS detector*, *Phys. Lett. B* **816** (2021) .

[78] CMS Collaboration, *Observation of Higgs boson decay to bottom quarks*, *Phys. Rev. Lett.* **121** (2018) , [1808.08242].

[79] CMS Collaboration, *Evidence for Higgs boson decay to a pair of muons*, *JHEP* **2101** (2020) , [2009.04363].

[80] ATLAS Collaboration collaboration, ATLAS Collaboration, *A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector*, *Phys. Lett. B* **812** (2021) .

[81] B. Carlson, T. Han and S. C. I. Leung, *Higgs boson to charm quark decay in vector boson fusion plus a photon*, *Phys. Rev. D* **104** (2021) , [2105.08738].

[82] G. T. Bodwin, F. Petriello, S. Stoynev and M. Velasco, *Higgs boson decays to quarkonia and the $H\bar{c}c$ coupling*, *Phys. Rev. D* **88** (2013) , [1306.5770].

[83] G. T. Bodwin, H. S. Chung, J.-H. Ee, J. Lee and F. Petriello, *Relativistic corrections to Higgs boson decays to quarkonia*, *Phys. Rev. D* **90** (2014) , [1407.6695].

[84] G. T. Bodwin, H. S. Chung, J.-H. Ee and J. Lee, *New approach to the resummation of logarithms in Higgs-boson decays to a vector quarkonium plus a photon*, *Phys. Rev. D* **95** (2017) , [1603.06793].

[85] C. Delaunay, T. Golling, G. Perez and Y. Soreq, *Enhanced Higgs boson coupling to charm pairs*, *Phys. Rev. D* **89** (2014) , [1310.7029].

[86] G. Perez, Y. Soreq, E. Stamou and K. Tobioka, *Constraining the charm Yukawa and Higgs-quark coupling universality*, *Phys. Rev. D* **92** (2015) , [1503.00290].

[87] G. Perez, Y. Soreq, E. Stamou and K. Tobioka, *Prospects for measuring the Higgs boson coupling to light quarks*, *Phys. Rev. D* **93** (2016) , [1505.06689].

[88] ATLAS Collaboration, *Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays to charm quarks with the ATLAS detector*, *Eur. Phys. J. C* **82** (2022) , [2201.11428].

[89] CMS Collaboration, *A search for the standard model higgs boson decaying to charm quarks*, *Journal of High Energy Physics* **2020** (Mar., 2020) .

[90] ATLAS Collaboration, *Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector*, .

[91] CMS collaboration, A. Tumasyan et al., *Search for Higgs Boson Decay to a Charm Quark-Antiquark Pair in Proton-Proton Collisions at s=13 TeV*, *Phys. Rev. Lett.* **131** (2023) , [2205.05550].

[92] Sherpa collaboration, Bothmann E. et al., *Event Generation with Sherpa 2.2*, *SciPost Phys.* **7** (2019) , [1905.09127].

[93] S. Catani, F. Krauss, R. Kuhn and B. R. Webber, *QCD matrix elements + parton showers*, *JHEP* **11** (2001) , [hep-ph/0109231].

[94] F. Krauss, *Matrix elements and parton showers in hadronic interactions*, *JHEP* **08** (2002) , [hep-ph/0205283].

[95] Hoeche S., Krauss F., Schumann S. and Siegert F., *QCD matrix elements and truncated showers*, *JHEP* **5** (2009) , [0903.1219].

[96] Winter J.-C., Krauss F. and Soff G., *A modified cluster-hadronisation model*, *The European Physical Journal C* **36** (2004) 381–395.

[97] Schönherr M. and Krauss F., *Soft Photon Radiation in Particle Decays in SHERPA*, *JHEP* **12** (2008) , [0810.5071].

[98] S. Badger, B. Biedermann, P. Uwer and V. Yundin, *NLO QCD corrections to multi-jet production at the LHC with a centre-of-mass energy of s = 8 tev*, *Physics Letters B* **718** (Jan., 2013) 965–978.

[99] Buckley A. et al., *Rivet user manual*, *Comput. Phys. Commun.* **184** (2013) 2803–2819, [1003.0694].

[100] Cacciari M. et al., *FastJet User Manual*, *Eur. Phys. J.* **C72** (2012) , [1111.6097].

[101] ATLAS Collaboration, *Secondary vertex finding for jet flavour identification with the ATLAS detector*, tech. rep., CERN, June, 2017.

[102] Abadi M. et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015.

[103] Shapley, L. S. , *A value for n-person games*, *Contributions to the Theory of Games (AM-28)* **2** (1952) 307–318.

[104] Hermann T., "Frugally Deep: Header-only library for using Keras models in C++.." https://github.com/Dobiasd/frugally-deep.

[105] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13\,TeV$*, *Eur. Phys. J. C* **79** (2019) .

[106] ATLAS Collaboration, *Performance and Calibration of the JetFitterCharm Algorithm for c-Jet Identification*, tech. rep., ATLAS Collaboration, 2015.

[107] Moneta L. et al., *The RooStats Project*, 2011.

[108] Verkerke W. and Kirkby D., *The RooFit toolkit for data modeling*, 2003.

[109] Brun R. and Rademakers F., *ROOT: An object oriented data analysis framework*, *Nucl. Instrum. Meth.* **A389** (1997) 81–86.

[110] ATLAS collaboration, ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, *Eur. Phys. J. C* **83** (2023) 982, [2212.09379].

[111] ATLAS collaboration, *Extrapolation of ATLAS sensitivity to $H \to b\bar{b}$ and $H \to c\bar{c}$ decays in VH production at the HL-LHC*, tech. rep., CERN, Geneva, 2021.

[112] EDF, "EDF Hartlepool power station website." `https://www.edfenergy.com/energy/power-stations/hartlepool`, 2023.

[113] EDF, "Nuclear generation: Our journey towards Zero Harm." `https://www.edfenergy.com/energy/safety-reporting`, 2023.

[114] G. C. R. George E. P. Box, Gwilym M. Jenkins and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Oxford University Press, 2015.

[115] I. Hughes and T. Hase, *Measurements and their uncertainties. A practical guide to modern error analysis* . July, 2010.

[116] P. Szendro, G. Vincze and A. Szasz, *Pink-noise behaviour of biosystems*, *European biophysics journal : EBJ* **30** (Aug., 2001) 227–31.

[117] R. F. Engle and C. W. J. Granger, *Co-integration and error correction: representation, estimation and testing*, *Econometrica* **55** (1987) 251–276.

[118] J. Aylett-Bullock, R. T. Gilman, I. Hall, D. Kennedy, E. S. Evers, A. Katta et al., *Epidemiological modelling in refugee and internally displaced people settlements: challenges and ways forward* , *BMJ Global Health* **7** (2022) .

[119] C. Altare, V. Kahi, M. Ngwa, A. Goldsmith, H. Hering, A. Burton et al., *Infectious disease epidemics in refugee camps: A retrospective analysis of UNHCR data (2009-2017)* , *Journal of Global Health Reports* **3** (2019) .

[120] T. Hoang, P. Coletti, A. Melegaro, J. Wallinga, C. G. Grijalva, J. W. Edmunds et al., *A Systematic Review of Social Contact Surveys to Inform Transmission Models of Close-contact Infections* , *Epidemiology* **30** (Sept., 2019) 723–736.

[121] K. van Zandvoort, M. O. Bobe, A. I. Hassan, M. I. Abdi, M. S. Ahmed, S. M. Soleman et al., *Social contacts and other risk factors for respiratory infections among internally displaced people in Somaliland*, *Epidemics* **41** (2022) .

[122] K. Prem, A. R. Cook and M. Jit, *Projecting social contact matrices in 152 countries using contact surveys and demographic data* , *PLOS Computational Biology* **13** (2017) .

[123] K. Prem, K. v. Zandvoort, P. Klepac, R. M. Eggo, N. G. Davies, C. f. t. M. M. o. I. D. C.-. W. Group et al., *Projecting contact matrices in 177 geographical regions: An update and comparison with empirical data for the COVID-19 era* , *PLOS Computational Biology* **17** (July, 2021) .

[124] D. Mistry, M. Litvinova, A. Pastore y Piontti, M. Chinazzi, L. Fumanelli, M. F. C. Gomes et al., *Inferring high-resolution human mixing patterns for disease modeling*, *Nature Communications* **12** (Dec., 2021) .

[125] S. Xia, J. Liu and W. Cheung, *Identifying the Relative Priorities of Subpopulations for Containing Infectious Disease Spread* , *PLoS ONE* **8** (June, 2013) .

[126] E. Zagheni, F. C. Billari, P. Manfredi, A. Melegaro, J. Mossong and W. J. Edmunds, *Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases* , *American Journal of Epidemiology* **168** (Nov., 2008) 1082–1090.

[127] F. Iozzi, F. Trusiano, M. Chinazzi, F. C. Billari, E. Zagheni, S. Merler et al., *Little Italy: An Agent-Based Approach to the Estimation of Contact Patterns-Fitting Predicted Matrices to Serological Data* , *PLoS Computational Biology* **6** (Dec., 2010) .

[128] I. Vernon, J. Owen, J. Aylett-Bullock, C. Cuesta-Lazaro, J. Frawley, A. Quera-Bofarull et al., *Bayesian emulation and history matching of JUNE*, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **380** (2022) .

[129] WHO, "Bangladesh - Rohingya Crisis: Early Warning, Alert and Response System (EWARS)."

https://www.who.int/bangladesh/emergencies/Rohingyacrisis/ewars, 2020.

[130] Government of the People's Republic of Bangladesh Office of the Refugee Relief and Repatriation Commissioner, " Rohingya refugee camp operations: Essential Programmes in light of COVID-19 ." http://rrrc.gov.bd/sites/default/files/files/rrrc.portal.gov.bd/notices/c3aece34_0550_4b4d_b33c_e8864272ada9/2020-03-25-16-34-21d19f130456961e35a25dbd1e5ef780.pdf, 2020.

[131] *Conversations with Cox's Bazar refugee operation Community Based Protection (CBP) team*, 2022.

[132] UNHCR, "The 10-Point Plan, Chapter 2: Data Collection and Analysis." https://www.unhcr.org/50a4c2b09.pdf, 2011.

[133] UNHCR, "Guidance on the protection of personal data of persons of concern." https://www.refworld.org/docid/5b360f4d4.html, 2018.

[134] UNHCR, WHO, UN Global Pulse, UN OCHA, Durham University, "Bangladesh: COVID-19 Exposure and Protective Measures." https://microdata.unhcr.org/index.php/catalog/587, 2020.

[135] P. Klepac, A. J. Kucharski, A. J. Conlan, S. Kissler, M. L. Tang, H. Fry et al., *Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project* , tech. rep., medRxiv, Mar., 2020. 10.1101/2020.02.16.20023754.

[136] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. No. 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[137] G. N. Lance and W. T. Williams, *Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses")*, *The Computer Journal* **9** (May, 1966) 60–64.

[138] S. Gupta, R. M. Anderson and R. M. May, *Networks of sexual contacts: implications for the pattern of spread of HIV*, *AIDS* **3** (1989) 807–818.

[139] C. P. Farrington, H. J. Whitaker, J. Wallinga and P. Manfredi, *Measures of disassortativeness and their application to directly transmitted infections.* , *Biom J.* **3** (2009) 387–407.

[140] C. Cuesta-Lazaro, A. Quera-Bofarull, J. Aylett-Bullock, B. N. Lawrence, K. Fong, M. Icaza-Lizaola et al., *Vaccinations or Non-Pharmaceutical Interventions: Safe Reopening of Schools in England*, *medRxiv* (2021) .

[141] Office for National Statistics, "Table ID KS105UK. (Household composition)." `https://www.nomisweb.co.uk/census/2011/KS105UK`, 2011.

[142] Office for National Statistics, " Ref: 008855. (Families with dependent children by number of children, UK, 1996 to 2017) ." `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families`, 2018.

[143] Office for National Statistics, "KS102UK. (Age Structure)." `https://www.nomisweb.co.uk/census/2011/KS102UK`, 2011.

[144] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, *Phys. Rev.* **145** (May, 1966) 1156–1163.

[145] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, *Phys. Rev. Lett.* **13** (Oct., 1964) 508–509.

[146] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, *Phys. Lett.* **12** (1964) 132–133.

[147] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, *Phys. Rev. Lett.* **13** (Aug., 1964) 321–323.

[148] G. S. Guralnik, C. R. Hagen and T. W. Kibble, *Global Conservation Laws and Massless Particles*, *Physics Review Letters* **13** (Nov., 1964) 585–587.

[149] C. A. Baker, D. D. Doyle, P. Geltenbort, K. Green, M. G. D. van der Grinten, P. G. Harris et al., *Improved Experimental Limit on the Electric Dipole Moment of the Neutron*, *Phys. Rev. Lett.* **97** (Sept., 2006) .

[150] V. Baluni, CP-*nonconserving effects in quantum chromodynamics*, *Phys. Rev. D* **19** (Apr., 1979) 2227–2230.

[151] Y. Fukuda, T. Hayakawa, E. Ichihara, K. Inoue, K. Ishihara, H. Ishino et al., *Measurement of the Flux and Zenith-Angle Distribution of Upward Throughgoing Muons by Super-Kamiokande* , *Physical Review Letters* **82** (Mar., 1999) 2644–2648.

[152] SNO Collaboration collaboration, Q. R. Ahmad, R. C. Allen, T. C. Andersen, J. D.Anglin, J. C. Barton, E. W. Beier et al., *Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory* , *Phys. Rev. Lett.* **89** (June, 2002) .

[153] K. Eguchi, S. Enomoto, K. Furuno, J. Goldman, H. Hanada, H. Ikeda et al., *First Results from KamLAND: Evidence for Reactor Antineutrino Disappearance* , *Physical Review Letters* **90** (Jan., 2003) .

[154] F. Zwicky, *Die Rotverschiebung von extragalaktischen Nebeln*, *Helv. Phys. Acta* **6** (1933) 110–127.

[155] V. C. Rubin and J. Ford, W. Kent, *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions*, *The Astrophysical Journal* **159** (Feb., 1970) .

[156] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro et al., *Measurements of* Ω *and* Λ *from 42 High-Redshift Supernovae*, *The Astrophysical Journal* **517** (June, 1999) 565–586, [astro-ph/9812133].

[157] A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich et al., *Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant*, *The Astrophysical Journal* **116** (Sept., 1998) 1009–1038, [astro-ph/9805201].

[158] L. Canetti, M. Drewes and M. Shaposhnikov, *Matter and antimatter in the universe*, *New Journal of Physics* **14** (Sept., 2012) .

[159] F. Feruglio, *Pieces of the flavour puzzle*, *European Physical Journal C* **75** (Aug., 2015) , [1503.04071].

[160] Kobo Inc., "KoboToolbox." https://www.kobotoolbox.org.

[161] Open Data Kit, "Open Data Kit." https://opendatakit.org.