A Comparison of Information Criteria for Traffic Model Selection

Paul Tune and Matthew Roughan
ARC Centre of Excellence for Mathematical and Statistical Frontiers
School of Mathematical Sciences
The University of Adelaide
Adelaide, Australia.
Email: {paul.tune,matthew.roughan}@adelaide.edu.au

Kenjiro Cho
IIJ Research Lab
Tokyo, Japan.
Email: kjc@iijlab.net

***Abstract*—Traffic modelling is a core component of network planning and engineering. Although good models are approximations of reality, they are very useful in various network applications. However, traffic modelling is often done in an *ad hoc* manner, guided only by the experience of the model designer. In this paper, we propose the use of *information criteria*, such as the Akaike Information Criterion (AIC), to systematically choose models. We study these criteria on Frequency, Frequency + Spike, and Wavelet models of the network traffic to select the best of these. However, there are many alternative information criteria, which give different results. We found that the Bayesian Information Criterion (BIC), and Minimum Description Length (MDL) provided better models than the (perhaps) more commonly used AIC and corrected AIC for network traffic modelling. Interestingly, we found that fancier models, such as Wavelet models, may reduce prediction accuracy, so simple frequency-based models are preferable.**

I. Introduction

Many network planning tasks require the traffic predictions as an input. There are various approaches one might apply to obtain such a prediction, but within each there is a model (albeit the model is sometimes implicit). There are a wealth of models for traffic (far too many to list here), so the question of which model is best arises. Typically, the question has been answered without clear scientific reasoning. For instance, models are often proposed as superior because they fit a dataset more accurately. This is a bad *sole* criteria.

Modelling requires a fine balance between simplicity and accuracy. It is obvious why the latter criteria is important, but the former requires some explanation. Simple models have many advantages:

- Simple models are often more tractable, *i.e.,* analysis of the model is easier;
- The parameters of simpler models may be measured from data more easily, and accurately; and
- Simple models are more *generalisable*, *i.e.,* the model estimated from one set of data is more likely to be helpful when applied to a new set of data.

The last property is critical for prediction.

We can always improve the fit of a model by increasing the size of the space of models. However, complex models often *overfit* a dataset. Even if such a model fits the current data more accurately, it likely provides poorer predictions. For instance, one could construct a model for a dataset that was just the data points themselves, but this would not typically provide better predictions than a model based on the "physics" of the problem, *i.e.,* the known rules that constrain the data. Hence the need to balance simplicity with the quality of fit to the data: the so-called *model selection problem.*

Information criteria (IC) were developed to address the model selection problem. IC have strong theoretical backing and guarantees, which can help model designers in their task to understand and systematically build suitable models of observed traffic data. IC do so by providing the optimal theoretical tradeoff between simplicity and accuracy of fit.

In this paper we apply IC to temporal modelling of network traffic by assessing three classes of models: Frequency (Fourier), Frequency + Spikes, and Wavelet models. To the best of our knowledge, *this is the first work to apply IC to network traffic modelling.* The main contribution of this paper is the methodology, which we advocate being applied to future modelling questions to balance "fit" and simplicity.

There is more than one information criterion. They differ in approach and assumptions about the data being studied. For instance, the Akaike information criterion (AIC) is based on asymptotic results, so it may not be optimal for finite datasets. We examine here five of the most commonly used criteria: the Akaike Information Criterion (AIC), the corrected AIC (AIC_c), the Bayesian Information Criterion (BIC), the two-stage Minimum Description Length (MDL) and normalised Minimum Description Length (nMDL). They produced a wide variety of models, and so the questions arise "how useful are these approaches if they are not consistent, and which should we prefer for traffic modelling?"

To answer these questions we applied the five IC to our real-world datasets, Abilene [24], GÉANT [31], and link traffic data from IIJ, an Internet Service Provider (ISP) in Japan (see §III), to obtain models of the traffic. These IC-derived models are then tested on their predictive ability: in particular, we use the IC to build predictive models of the traffic and then evaluate their predictions.

The results are clear: AIC and AIC_c tend to overfit the data. At best they provide results on par with the other approaches

(with the cost of many more parameters), but typically they are worse. The performance of BIC and (n)MDL are roughly the same, though there are differences in some settings.

The results, in general, suggest that the traffic should be modelled with the Frequency + Spike model as it provides a good tradeoff between modelling error and model complexity. In comparison, fancier models based on wavelets actually reduce forecast accuracy.

II. Information Criteria

Given a dataset, the first obvious decision by a model designer is the choice of the class of model. There are many different models to choose from.

Another important consideration, often overlooked, is the choice of the number of parameters of the model, or the *model order*. Often equated with complexity, the number of parameters is a crucial choice. If we allow more, we allow a larger space of models, and so must fit our data better. This may seem preferable, but too many parameters may result in over-fitting, as well as other problems. However, too few parameters and we may miss out on legitimate features of the data.

The first breakthrough in a systematic method for model selection was an information criterion proposed by Akaike [3]. Known as the Akaike Information Criterion (AIC), it is

$$\text{AIC} = 2p - 2\log L(\theta^*), \tag{1}$$

where p is the number of parameters and $L(\theta^*)$ is the maximised value of likelihood of the model using the optimal choice of parameter θ^*. If the parameter θ^* is unknown, it is often substituted by the maximum likelihood estimate of θ^*. We can see the trade-off between the complexity of the model, measured by the first term in (1) and the error of the fit, measured by the second term. To select the best model, one selects the candidate model with the minimum AIC value.

The beauty of the AIC is its simplicity, but despite this, it comes with a strong theoretical justification. The AIC is grounded in information theory. It quantifies the *information loss* when the true model of the data is not selected.

The information loss is measured by the Kullback-Leibler (KL) divergence $D_{\text{KL}}(f \,||\, g)$ between the observed process f and model g [8]. If an unknown process f can be represented by models g_1 or g_2, the criterion computes $D_{\text{KL}}(f \,||\, g_i)$ and chooses the model with the lower value, since this model minimised information loss.

Of course, if we already knew f, we can choose the correct model. Akaike's contribution was to show how the quantity $D_{\text{KL}}(f \,||\, g)$ can be *estimated* for any candidate model g without prior knowledge of f. When applied to a set of models $\mathcal{M}$, the AIC guarantees the selection of the model $g^* \in \mathcal{M}$ closest to f in KL divergence [3]. Additionally, it is an asymptotically efficient criterion for model selection [16].

There is a notable drawback. The AIC was derived under the assumption that n, the number of data points, is very large. This results in the AIC consistently overfitting for small n. Examples of AIC selecting higher model orders than necessary can be found in [7, Section 8.3]. The corrected AIC (AIC_c) [16] was developed to address this problem

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2p(2p+1)}{n-p-1}. \tag{2}$$

This criterion converges to the AIC as n increases, but outperforms the AIC on finite data.

The Bayesian Information Criterion (BIC) was developed by Schwarz [29] from the perspective of Bayesian statistics. It is also an asymptotic result assuming that the observed data's distribution belongs to the exponential family,

$$\text{BIC} = 2p\log n - 2\log L(\theta^*). \tag{3}$$

The BIC is not necessarily asymptotically efficient (unlike AIC and AIC_c [16]), but it is a consistent criterion. If the observed process is an observation of a model g, then as $n \to \infty$ BIC will select the model g with probability 1. Observe that the BIC has a larger penalty on the number of parameters chosen – BIC will tend to choose models with fewer parameters than AIC or AIC_c.

Finally, a different take on information criteria based on coding theory was introduced by Risannen [27]. The approach, known as the Minimum Description Length (MDL) criterion, can be summed up as exploiting the regularity of the data to compress it, or equivalently, to describe it in the minimum number of symbols (typically in bits) [14]. MDL is not concerned with the actual encoding of the data, but only the length of the encoding.

The original formulation of MDL involved a two-stage coding scheme: the overall description length of a model is the sum of the number of bits to describe the model (including the number of parameters) and the description of the data when encoded by the chosen model. Formally,

$$\text{MDL} = \mathcal{L}(g) + \mathcal{L}(f \,|\, g), \tag{4}$$

where g is the candidate model, f is set of data and $\mathcal{L}(\cdot)$ computes the length of the encoding. The two-stage MDL is closely related to the BIC; in a number of instances, both criteria select exactly the same model.

The two-stage MDL employs a particular *universal code* to encode the model and data. An alternative universal code is the normalised maximum likelihood code, leading to a different criterion called the normalised MDL (nMDL). Unlike two-stage coding, the parameters and model are jointly coded, based on maximum likelihood coding by Shtarkov [15]. This has the effect of reducing the number of codewords considered to encode the model g, thereby producing a more compact description length [14], [15] compared to two-stage MDL. There is no closed form for nMDL in general, but useful asymptotic expressions can be derived (see §III-A).

There are other approaches to model selection [7], but in this paper we focus on these, the most common approaches. We aim to use them on the problem of traffic modelling, but the differences between them that we shall see in the following section lead to a question about which is most useful, which we shall address in the subsequent section.

III. Model Selection in Action

In this section, we formulate our traffic model. We discuss the three dictionary functions used here, namely the Frequency, Frequency + Spike and Wavelet dictionaries. We also apply five IC to real-world datasets to test their performance.

A. Traffic Model

Here we primarily consider temporal models of traffic volume (though the concept of IC is much more general). It is well-known that network traffic has a strong diurnal and weekly component, giving rise to an almost periodic structure, with some (normal) stochastic variations, and spurious (spiky) traffic [13], [17], [19], [28].

There are many approaches to modelling the temporal nature of the traffic [33]. Examples include the Holt-Winters smoothing approach [4], stochastic modelling techniques such as the Norros model [28] and cyclo-stationary models [30]. Since we aim to predict traffic, we explore three classes of temporal models: one based on Fourier (Frequency) coefficients, one on Frequency + Spikes, and one on Wavelets. There are many other possibilities, but we *a priori* know that there is a strong diurnal periodic component to the signal [13], [19], [28], and the approaches above match this well.

Let $\boldsymbol{\theta}$ be a set of the model parameters of a model class $\mathcal{M}$, for instance the amplitude and phase of each frequency component in a simple Fourier model. Naïvely, with this model of a signal we would have as many parameters (the amplitudes and phases) as in the original signal. However, while many of these frequencies are needed to *reconstruct* the original data, many do not describe general properties of the signal so much as specific local variations that are not useful for predicting future traffic. So not all the parameters in $\boldsymbol{\theta}$ are needed when a model from the class $\mathcal{M}$ is selected. The number we do need is the *order* of the model within that class. IC are used to choose the order of the model.

In order to apply the IC, we need to choose the class of models. We take the class described by a linear combination of m *dictionary* functions $\mathbf{d}_i$, plus some *noise*,

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{z}, \tag{5}$$

where $\mathbf{D} = [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_m]$, and $\mathbf{z}$ denotes the *independent and identically distributed* (IID) Gaussian noise vector with mean 0 and variance σ^2. The reason we used a simple noise process is that we want the structured content of the signal to appear in the model, not the noise.

The $\mathbf{x}$ are called *predictors*, and combined with the noise variance, they form the model parameters $\boldsymbol{\theta} = \{\mathbf{x}, \sigma^2\}$, so the total number of parameters is $p = m + 1$.

The problem might seem almost trivial – surely we can just determine $\mathbf{x}$ by linear regression on $\mathbf{y}$? The difficulty is that we must also determine the matrix $\mathbf{D}$, *i.e.,* we must also select the particular dictionary functions we will use from a larger set. If the set of dictionary functions forms a basis, then there are simple algorithms for obtaining (5) for any particular $m < n$. However, in some cases we start with an *overcomplete* set in the hope of obtaining a sparser representation of data compared to a decomposition into basis vectors [6].

B. Choice of Dictionary

We start with the *Fourier-frequency model*, motivated by the strong diurnal pattern in traffic [28]. Here the vectors $\mathbf{d}_i$ are sinusoidal functions. These form a basis, and the algorithm for efficiently transforming a discrete signal into this basis is commonly called the Fast Fourier Transform (FFT). The FFT will transform a real-numbered dataset of length n into $n/2$ complex numbers (since the signals are real).

Since we are dealing with finite length segments of a (periodic) signal, the FFT sees discontinuities at the edges and unwanted frequencies are introduced *i.e.,* suppress spectral leakage. A windowing function is first applied to the data to minimise this effect. We have tested several windows, such as the Hanning and Blackman windows. Ultimately, we settled on the Hamming window since the performance with the other windows are similar, but unlike the other two windows, we can invert the Hamming window to obtain an accurate computation of the model error (see below).

We determine a representation (5) for $m < n$ simply by *forward selection* based on the magnitude of these complex coefficients. Coefficients are selected in descending order.

The Fourier transform of the data, however, loses all time-based activity in favour of the frequency-based view. Traffic also has "spikes" [19], for instance, as a result of anomalies. A spike (in time) is wideband in the Fourier domain, which means it is not well-represented by a small number of Fourier coefficients. It therefore stands to reason that a simple Fourier analysis of the data will have to compensate for these spikes with additional frequency coefficients.

It might be more efficient to consider a dictionary that includes both Fourier components, and spikes (formally a spike is just a δ-function in time). We call the dictionary the *Frequency + Spike model*. The set of dictionary functions is overcomplete, and so decomposing the signal into this basis requires a more sophisticated algorithm. Luckily, there is a now well-developed area of signal processing referred to as *compressive sensing* [5], [12] whose strength lies in finding sparse sets of dictionary functions to represent datasets.

Coefficients are selected by using the IC combined with Orthogonal Matching Pursuit (OMP) [11], [25], a popular greedy algorithm used for selecting $\mathbf{d}_i$ from overcomplete dictionaries. The selection procedure, however, is suboptimal in that not all subsets are searched. OMP selects a subset of $\mathbf{d}_i$s with the highest correlation to the signal. While computationally tractable, the selected coefficients are not guaranteed to be optimal. Despite this, we have obtained useful results.

The third set of dictionary functions we consider is based on the wavelet transform, which like the Fourier transform, is a transform from a time-based representation of the data, into the dictionary space, but the wavelet basis is a time-frequency representation. Each *wavelet* is like taking a particular frequency band, and localising it in time [22]. The traffic, while smooth, is only partially periodic with many transient

Dataset	Duration (weeks)	Start Date	Measurement interval	n	$\bar{n}$
Abilene	2	1 Mar 2004	5 minutes	4032	4258
GÉANT	6	1 Jan 2005	15 minutes	4032	4258
Japan 1	19	3 Apr 2014	1 hour	3192	3254
Japan 2	4	3 Apr 2014	1 hour	672	735

TABLE I
SUMMARY OF THE DATASETS. n AND $\bar{n}$ DENOTES THE NUMBER OF DATA POINTS AND THE PADDED LENGTH FOR WAVELET TRANSFORMS RESPECTIVELY.

frequency components. Wavelets should excel in situations like this, as the right choice of wavelet can localise frequencies well enough to provide a succinct representation of the traffic. Here, we limit ourselves to orthogonal wavelet representations, so the dictionary set is a basis, and the transform is invertible. The number of potential parameters may, however, be a little longer than n due to padding for computational efficiency purposes. In the results for wavelets below, we list this augmented length as $\bar{n}$.

The discrete wavelet transform is computed via filter banks [32]. In particular we used the function `wavedec` from the Wavelet Toolkit in `Matlab` to decompose the signal into wavelets, and `waverec` to reconstruct the model.

Wavelet coefficients are selected in the same manner as the Frequency dictionary *i.e.,* via forward selection.

We tested the Haar wavelet [22, p. 248], Daubechies wavelet with 6 vanishing moments, denoted by *db6* [22, p. 249], and the symlet with 8 vanishing moments, denoted by *sym8* [22, p. 253]. We only present the results for the *db6* wavelet due to space constraints, however, the Haar wavelet performed the worst, since it is non-smooth and applying it to a relatively smooth dataset produced a suboptimal representation, while the *sym8* wavelet performs as well as the *db6* wavelet.

C. Information Criteria

Let p denote the total number of parameters of the model. Let K, L and M the total number of frequency, spike and wavelet coefficients respectively. For a real signal, frequency components come in pairs, so we count this as specifying $2K$ parameters. Furthermore, a spike component requires specification of the location and magnitude, so this counts as $2L$. For wavelets, the number of coefficients are equal to the total number of predictors. Thus, the number of parameters for each dictionary is computed as:

- *Frequency*: $p = m + 1 = 2K + 1$,
- *Frequency + Spike*: $p = m + 1 = 2K + 2L + 1$,
- *Wavelet*: $p = m + 1 = M + 1$.

Let the *Residual Sum of Squares* be denoted by

$$\text{RSS} = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2, \tag{6}$$

which is the squared difference between the model's prediction, and the actual data. The RSS is a measure of the accuracy of the model's fit to the dataset.

In the context of the model class described above, there are tractable formulae for each of the IC:

$$\begin{aligned}
\text{AIC} &= 2p + n(\log \text{RSS} - \log n + 1),\\
\text{AIC}_\text{c} &= \text{AIC} + \frac{2p(p+1)}{n-p-1},\\
\text{BIC} &= p\log n + n(\log \text{RSS} - \log n + 1),\\
\text{MDL} &= \frac{p}{2}\log n + \frac{n}{2}\log \text{RSS},\\
\text{nMDL} &\approx \frac{p}{2}\log(\|\mathbf{y}\|_2^2 - \text{RSS}) - \frac{n-p-1}{2}\log(n-p)\\
&\quad -\frac{p+3}{2}\log p + \frac{n-p}{2}\log \text{RSS}.
\end{aligned}$$

Here, n is fixed, and so the terms involving only n could effectively be dropped in calculations of the optimal p. That means that the two-stage MDL is equivalent to the BIC so we omit the results for MDL. There is no closed form for the nMDL. The form presented above is asymptotic approximation given in [15], which we use in our experiments.

Each case has a term proportional to $\log \text{RSS}$, which measures the model's fit quality, with a penalty term that increases with the number of parameters p. The goal is to balance these two factors, but each IC does so differently.

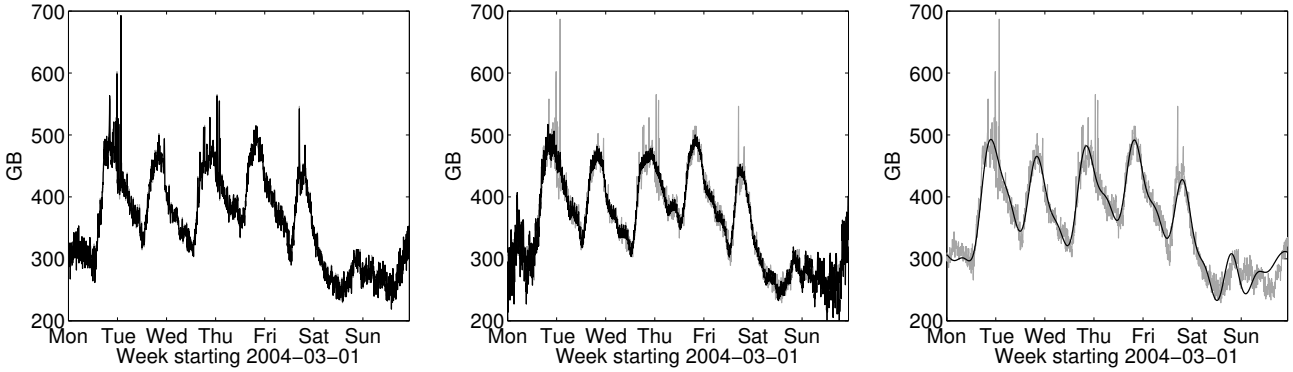
The IC themselves are *general*: as long as there are potential models of the traffic, then the IC can be used in conjunction with the models to select the appropriate model. They are general enough to be applied to mixture models *i.e.,* a convex combination of models (see [1] for the application of BIC to mixture models for modelling endhost traffic).

In practice, we can widen the class $\mathcal{M}$ to include models beyond the linear model used here, then apply IC to select the model most appropriate for describing the data. This does not apply only to parametric models; IC can be used with non-parametric models as well. The best model could conceivably be a convex combination of parametric and non-parametric models, but is guaranteed to have the best tradeoff between fit and complexity under the particular IC.

D. Results

Our experiments used 4 datasets summarised in Table I. The Abilene [24] and GÉANT [31] datasets come from the Abilene and GÉANT network based in North America and Europe respectively. The Japan 1 and 2 datasets come from IIJ, a Japanese ISP.

Preprocessing of the raw data was required. Traffic data may contain very large abrupt changes, so there are intervals in the datasets where our model cannot be applied to. For instance, in the Japan 1 and 2 datasets, we have observed large stepwise shifts in traffic due to traffic rerouting. In order to select contiguous measurement intervals in the dataset that are "stable" enough, *i.e.,* the time average has no abrupt shifts, we apply a *step detection* method to our datasets. We tried several step detection methods, but settled on a jump penalty detection algorithm [21], based on the Potts model, because fast implementations are available [23]. Using this on the raw datasets, we extracted our 4 datasets. The limitation of



(a) AIC (b) BIC (c) nMDL

Fig. 1. Comparison of AIC, BIC and nMDL on the Frequency model on 5 minute intervals of Abilene data for a single week starting 1st March 2004. AIC, BIC and nMDL selected models with 837, 32 and 11 frequencies respectively. The light grey curve is the actual traffic, while the dark curves are the models. We can see that AIC tries to fit all the datapoints, while nMDL tries to find the underlying pattern.

IC	Frequency			Frequency + Spike				Wavelet *db6*		
	Components	p	$\sqrt{\text{RSS}}$	Frequency	Spikes	p	$\sqrt{\text{RSS}}$	Components	p	$\sqrt{\text{RSS}}$
AIC	837	1675	296.2	20	12	52	570.4	2077	2078	0.0
AIC_c	107	215	736.2	20	12	52	570.4	870	871	165.9
BIC	32	65	883.3	18	8	44	608.5	210	211	415.0
nMDL	11	23	970.2	20	12	52	570.4	247	248	388.0

TABLE II
ABILENE DATA MODEL COMPARISON (5 MINUTE INTERVALS, 1 WEEK).

IC	Frequency			Frequency + Spike				Wavelet *db6*		
	Components	p	$\sqrt{\text{RSS}}$	Frequency	Spikes	p	$\sqrt{\text{RSS}}$	Components	p	$\sqrt{\text{RSS}}$
AIC	11	23	501.3	13	3	32	327.0	733	734	0.0
AIC_c	11	23	501.3	13	3	32	327.0	327	328	90.7
BIC	11	23	501.3	12	2	28	353.6	113	114	220.1
nMDL	11	23	501.3	13	3	32	327.0	129	130	203.1

TABLE III
ABILENE AGGREGATED DATA MODEL COMPARISON (15 MINUTE INTERVALS, 1 WEEK).

IC	Frequency			Frequency + Spike				Wavelet *db6*		
	Components	p	$\sqrt{\text{RSS}}$	Frequency	Spikes	p	$\sqrt{\text{RSS}}$	Components	p	$\sqrt{\text{RSS}}$
AIC	798	1597	142.2	55	13	137	496.8	3254	3255	0.0
AIC_c	527	1055	208.0	55	13	137	496.8	3191	3192	1.1
BIC	133	267	399.8	28	0	57	600.2	3191	3192	1.1
nMDL	96	193	442.8	44	3	92	535.8	1115	1116	374.0

TABLE IV
JAPAN 1 DATA MODEL COMPARISON (HOURLY INTERVALS, 19 WEEKS).

IC	Frequency			Frequency + Spike				Wavelet *db6*		
	Components	p	$\sqrt{\text{RSS}}$	Frequency	Spikes	p	$\sqrt{\text{RSS}}$	Components	p	$\sqrt{\text{RSS}}$
AIC	246	493	5.6	20	14	55	29.2	566	567	0.0
AIC_c	41	83	32.7	14	5	34	34.5	503	504	1.2
BIC	21	43	41.5	10	3	27	37.7	503	504	1.2
nMDL	21	43	41.5	13	5	37	34.6	188	189	443.0

TABLE V
JAPAN 2 DATA MODEL COMPARISON (HOURLY INTERVALS, 3 WEEKS).

this method is that our model is less descriptive than one that incorporates an abrupt shift component into the model. However, our motivation is to study the IC themselves, using a simple model with tractable expressions of the IC is the main focus here.

It is important that the prediction tests be performed on data that is *not* used in calculating the model, so we will fit our model only to a segment of the data, and preserve the rest for use in testing. The testing periods are the last single week of the datasets.

IC	Frequency			Frequency + Spike				Wavelet *db6*		
	Components	p	$\sqrt{\text{RSS}}$	**Frequency**	**Spikes**	p	$\sqrt{\text{RSS}}$	**Components**	p	$\sqrt{\text{RSS}}$
AIC	2015	4031	1.0	100	34	269	13.4	4030	4031	0.0
AIC_c	87	175	36.6	100	34	269	13.4	1743	1744	3.0
BIC	85	171	36.7	74	24	197	16.6	649	650	7.3
nMDL	85	171	36.7	74	24	197	16.6	675	676	7.1

TABLE VI
GÉANT DATA MODEL COMPARISON (15 MINUTE INTERVALS, 6 WEEKS).

Our objective is to fit a sparse linear model per equation (5) based on a particular dictionary to an observed sequence of traffic data. We take the common step in all cases to remove the mean of the traffic which is a standard preliminary.

Tables II to VI list the number of components chosen by each IC on the datasets, and the resulting number of parameters in the model. They also list the square root of the RSS, which shows the inverse relationship between the number of parameters and model accuracy clearly. The more parameters the smaller the RSS, though the exact relationship varies by model. Note that the RSS can only be compared within, but not between, datasets because of the different units of traffic involved.

Naïvely, we would prefer models with a good fit *i.e.,* low RSS, but the various IC attempt to perform a tradeoff between the number of parameters p and the RSS. We see that AIC consistently chooses a tradeoff with a large number of parameters. The literature [7, Section 8.3] suggested that this would be a result of using a finite dataset, but although AIC_c attempts to correct this, the number of parameters included is still large. Both BIC and nMDL chooses substantially fewer parameters (as earlier noted MDL returns the same results as BIC). The obvious question then is which is most useful. We shall aim to answer that in the next section.

Figure 1 shows three examples of the models chosen by the IC for the Frequency dictionary. Comparing the figures, BIC and nMDL chose a model with far less frequencies compared to AIC's, so their models are smoother, though this is less apparent for BIC. However, in all cases, a small number of components chosen leads to a model that can't recreate the (wide-band) spikes. Hence the Frequency + Spike model.

The other main comparison is between models – which of these is actually better. If, for instance, we were to choose BIC or nMDL as our preferred criteria (we shall see why this might be the case in the following section), then we might compare the IC values between models. For our datasets, the Frequency + Spike model is the most preferable (with the exception of nMDL). This seems to be the model that provides the best tradeoff between representing the data with the fewest parameters and fit error.

There is a codicil on the Frequency + Spike model. We aim to use it in prediction, but we also know that spikes potentially originate from anomalous traffic – so they may not generalise. We shall consider two approaches, the Frequency + Spike model itself, and a modified model where the spikes are used in modelling the traffic, but excluded in prediction. That way they can still be used to diffuse some of the wide-band energy, without compromising predictions.

We repeated these experiments for a range of cases, and present here in Table III the results when the data is further aggregated into time intervals of 15 minutes. The base number of data points is $n = 672$, so naturally the models have fewer parameters as a result of the smoothing of the data resulting from aggregation.

We immediately see a reduction in coefficients for all criteria. Moreover, for the Frequency model, all criteria are picking the same coefficients. The aggregation of traffic in longer measurement intervals has smoothed out spurious frequencies, acting as a low pass filter. Interestingly, nMDL's selection remains unchanged, demonstrating its robustness.

The number of coefficients selected, however, does not proportionally decrease for each criterion. The reduction in parameters for AIC and AIC_c are large, as would be expected given the large number of parameters these approaches select. The reduction in BIC and nMDL parameters are much smaller. Intuitively, these IC are already producing "smooth" models, and the models therefore don't change much when we aggregate/smooth the input data. This is an indication of generalisability of the models, but we shall see clearer evidence in the following section.

Finally, we clearly see that the Wavelet models perform poorly on the datasets. In particular, for the Japan datasets where the periodicity is more pronounced, almost all coefficients are selected by most of the IC, including BIC. The nMDL criterion here still outperforms the other criteria by selecting far less coefficients than the competition.

E. Frequency analysis

By the strong cyclical behaviour of the network traffic we have seen, in this section, we study the Fourier transforms of the datasets (via FFT) to find the prominent frequencies and their harmonics. Most traffic models, as far as we know, do not account for harmonics in the traffic data. Exploiting knowledge of the harmonics can lead to a more compact model, since the number Fourier basis vectors, for instance, is reduced.

In our data, we expect to find hourly, daily and weekly cycles. These correspond to 1, 1/24 and 1/168 cycles per hour respectively. There may be, perhaps, other frequencies too. Note that the daily cycle's harmonics may coincide with the hourly cycle's harmonics, and similarly with the weekly and daily cycles. We therefore only count the harmonics belonging to the larger cycle, *e.g.,* a weekly cycle's harmonic that aligns with a daily cycle's is counted as the daily cycle's harmonic.

Criterion	Hourly	Daily	Weekly	Other
Abilene	2	4	3	0
AIC	0	1	2	0
AIC_c	0	1	2	0
BIC	0	1	2	0
nMDL	0	1	2	0

TABLE VII
COMPARISON OF PROMINENT FREQUENCIES AND THEIR HARMONICS FROM ACTUAL DATA AGAINST THE MODELS SELECTED BY THE IC ON ABILENE DATA (15 MINUTE INTERVALS, 1 WEEK).

Criterion	Daily	Weekly	Other
Japan 1	7	11	65
AIC	7	11	65
AIC_c	7	11	65
BIC	5	8	56
nMDL	3	8	53

TABLE VIII
COMPARISON OF PROMINENT FREQUENCIES AND THEIR HARMONICS FROM ACTUAL DATA AGAINST THE MODELS SELECTED BY THE IC ON JAPAN 1 DATA (HOURLY INTERVALS, 19 WEEKS). NOTE THAT HOURLY FREQUENCY CYCLES ARE UNOBSERVABLE AS THE MEASUREMENT INTERVALS ARE HOURLY.

We then compare against the Frequency models selected by the IC to see if the prominent frequencies were selected. The dataset is preprocessed by applying a Hamming window on it. We then set (by thresholding) all frequencies below the smallest magnitude weekly harmonic we could find to 0. For Abilene, this would be 25th harmonic, and in the 19 week Japan 1 data, the 84th harmonic. The same threshold is also applied to the models selected by the IC, so the results presented here may have a total number of frequencies less than the results in previous section.

In the results below, the number of weekly harmonics present is larger than the number of daily (or hourly) harmonics. Intuitively, the weekly cycle has far more complex structure than the hourly and daily cycles.

Table VII presents the results for the Abilene data with 15 minute measurement intervals for a week. Since the measurement intervals are less than an hour, we expect to find prominent hourly, daily and weekly frequencies. We find that none of the selected models picked out the hourly frequencies, instead picking just the daily and weekly frequencies. It turns out that the latter two frequencies are also the largest.

Compared to Table III, it seems like the selected models have included a lot more frequency components than necessary. On closer inspection, many of these components are additional frequencies that appear due to the quantisation effect of FFT. Since FFT is effectively a discrete Fourier transform (DFT), the additional frequencies arise because the data's prominent frequencies do not perfectly align with the quantised bins of the DFT. Though we have performed most suppression of these frequencies via preprocessing with the Hamming window, another way to get around this effect is to include the traffic sampling effects into the model (5). The new model, however, would be more complicated (and changes the form of the IC themselves). We leave this for future work.

We then examined the Japan 1 data with the results in Table VIII. Here, due to the hourly measurement intervals, hourly cycles cannot be observed, because its frequency is higher than the measurement (sampling) rate. We find that there are more harmonics present compared to the Abilene data. Compared to the actual data, the selected Frequency models by the IC are picking most of the daily and weekly harmonics. AIC and $AIC_c$are selecting all the harmonics, while BIC and nMDL are more conservative in their selection.

While some of the Other frequencies are due to FFT quantisation effects, the majority of the Other frequencies in the Japan dataset are mostly due to the harmonics arising from a finite length sample. These are multiples of 1/3192 cycles per hour, since our data length is 19 weeks. BIC and nMDL demonstrate their robustness by selecting less of these Other frequencies, unlike AIC and AIC_c.

IV. PREDICTIVE POWER EVALUATION

Here, we evaluate the predictive power of the IC with the three dictionaries. When comparing the IC selected models, it is not enough to compare them on their fit; the complexity of the model must also be included. The fit is computed by the RSS and the complexity of the model is determined by p in the previous section.

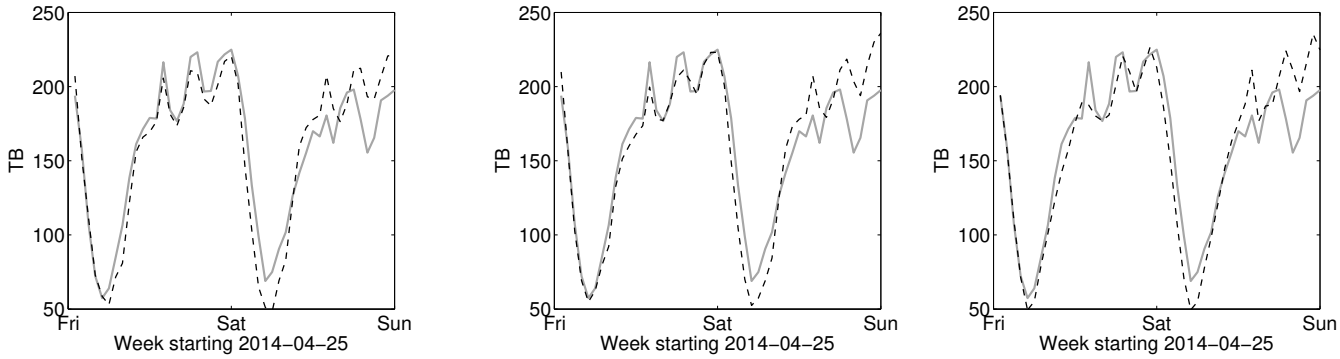
We apply the IC under the models from the last section to historical data and then obtain a prediction for a single week. Note that the week-long traffic to test the models' predictive power is not part of the historical data used to train the models.

Figure 2 shows the predictions made by the Frequency, Frequency + Spike and Wavelets models, via nMDL for a single week from the Japan 1 data, with 48 hours plotted. All models generally fit the diurnal cycles of the weekly traffic quite closely, as all IC select the most important frequencies of the diurnal cycle. What differs is their complexity; the Wavelet models are the most complex as they contain the most number of coefficients. The Frequency + Spike model (with no Spikes) suffices in this case, since it has a low number of coefficients with RSS competitive to the Frequency model.

For a clearer comparison, Tables IX to XII presents the RSS error of the predictions made by the models for the four IC.

In all cases, the Frequency + Spike model has the best predictive power to the number of coefficients, but predictions with spikes do not necessarily give better performance. Spikes are often due to anomalies, for instance, denial of service attacks, and flash crowds [28]. When these anomalies will occur is difficult to determine. Error is reduced in some cases with spikes (*e.g.,* Table IX) but this is not always the clear winner. So spikes do not really aid in prediction, and should be omitted. For instance, Figure 3 shows an example on Abliene data where spikes were predicted, but the actual traffic do not contain these spikes, resulting in significant prediction error.

There is little difference between predictions under various IC because only one model class is used as per (5). For example, in [10], several model classes were used and evaluated together with various IC. The chosen models has a more significant difference in performance as there is more



(a) Frequency (b) Frequency + Spike (no spike) (c) Wavelet

Fig. 2. Predictions via the nMDL using the Frequency, Frequency + Spike (with no spikes) and Wavelet models on hourly intervals of Japan 1 data on a single week starting 25th April 2014. For clarity, an interval of 48 hours is plotted. The light grey curve is the actual traffic, while the dashed curves are the predictions. The close fit for all models is due to the models selecting the main frequencies that captures the diurnal cycles of weekly traffic. Here, the Frequency model has the lowest RSS error.

IC	Frequency	Frequency + Spike		Wavelet *db6*
		No Spikes	With Spikes	
AIC	1437.9	1375.4	1254.4	1405.0
AICc	1431.3	1375.4	1254.4	1407.7
BIC	1395.1	1357.9	1256.1	1390.6
nMDL	1436.9	1375.4	1254.4	1394.2

TABLE IX
RSS FOR ABILENE AGGREGATED DATA MODEL COMPARISON. PREDICTIONS WERE MADE USING DATA A WEEK PRIOR WITH 15 MINUTE INTERVALS.

IC	Frequency	Frequency + Spike		Wavelet *db6*
		No Spikes	With Spikes	
AIC	196.6	195.7	202.4	196.1
AICc	191.3	195.7	202.4	196.2
BIC	175.5	195.7	202.4	196.2
nMDL	174.1	195.7	202.4	202.9

TABLE X
RSS FOR JAPAN 1 DATA MODEL COMPARISON. PREDICTIONS WERE MADE USING DATA 19 WEEKS PRIOR WITH HOURLY INTERVALS.

IC	Frequency	Frequency + Spike		Wavelet *db6*
		No Spikes	With Spikes	
AIC	244.6	240.9	242.8	244.6
AICc	244.2	240.7	243.0	244.6
BIC	244.4	241.4	243.2	244.6
nMDL	244.4	240.8	243.1	244.3

TABLE XI
RSS FOR JAPAN 2 DATA MODEL COMPARISON. PREDICTIONS WERE MADE USING DATA 3 WEEKS PRIOR WITH HOURLY INTERVALS.

IC	Frequency	Frequency + Spike		Wavelet *db6*
		No Spikes	With Spikes	
AIC	37.9	37.1	37.0	37.9
AICc	38.8	37.1	37.0	37.9
BIC	38.4	36.7	36.5	37.8
nMDL	38.4	36.7	36.5	37.8

TABLE XII
RSS FOR THE GÉANT DATA MODEL COMPARISON. PREDICTIONS WERE MADE USING DATA 6 WEEKS PRIOR WITH 15 MINUTE INTERVALS.

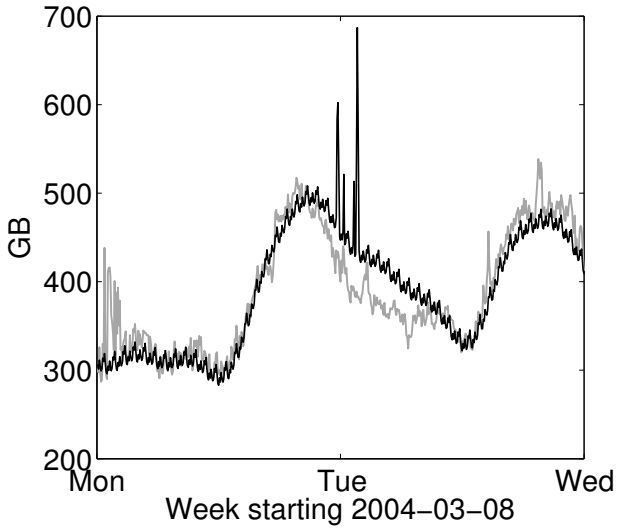


Fig. 3. Prediction via the nMDL for the Frequency + Spike (with spikes) on 5 minute intervals of Abilene data on a single week starting 8th March 2004. For clarity, an interval of 48 hours is plotted. The light grey curve is the actual traffic, while the black curve is the prediction. We see here that the predicted spikes are not present in the actual traffic, causing significant prediction error.

flexibility in choice. Significant differences are likely if the model class is widened from just (5), for instance, allowing a trending and jump component to better fit increases in traffic and level shifts in traffic due to route changes.

Wavelet models perform only as well, or slightly even worse than Frequency models. In most cases, the models selected are the same under different IC, so the resulting RSS is the same. This shows that complex, fancier models do not necessarily perform better, but may be detrimental to prediction.

Between all IC, nMDL often performs better, with BIC coming in at a close second. The results show that the models with lower complexity are better predictors of future traffic than their competitors.

V. Related Work

There are many traffic models, most focusing on modelling the heavy tails of network traffic [9], [20], [26]. Due to limited space, we cover the most relevant ones here. Roughan *et al.* [28] described a temporal model of traffic flows on the backbone network that accounts for seasonal traffic variations and long term growth. Lakhina *et al.* [19] showed how principal component analysis (PCA) can be used to extract features about the traffic. In particular, they showed three distinct types of Origin-Destination network flows: diurnal, spurious and noisy flows. These insights can be used to detect anomalies [18].

Although IC are a standard practice of stochastic modelling, to the best of our knowledge, few works have applied IC to Internet traffic modelling. Agosta *et al.* [1], for instance, used the BIC to approximate the Bayes factor when selecting mixture models to model endhost traffic. Tenório et al. [10] studied the performance of various IC on a model used for blind detection of malicious traffic and concluded that the Efficient Determination Criterion and the Exponential Fitting Test performed the best for their application.

In terms of applying overcomplete dictionaries to network traffic, Aiello *et al.* [2] tested three overcomplete dictionaries for compression: the Fourier + Haar wavelets, Fourier + Spike and Fourier + Spike + Haar wavelets, with selection performed by a greedy pursuit algorithm. They noted that compression is better with a larger overcomplete dictionary but at the cost of higher computational complexity.

Their focus, however, was on data compression, so there was an emphasis on high reconstruction fidelity, whereas here we care about modelling and prediction. Their insights, however, suggest the Fourier + Spike dictionary will outperform the other two, which is indeed the case here as well.

Our work aims to systematise Internet traffic modelling by introducing IC as a formal procedure that supplements modelling and provides theoretical grounding to model selection. For instance, IC can be used together with PCA [19] to select a model about the network traffic flows.

VI. Conclusion

In this paper, we studied the application of information criteria for selecting appropriate models for network traffic. We compared five information criteria: AIC, AIC_c, BIC, two-stage MDL and nMDL. The crucial insight of these criteria is that complex models can over-fit a dataset, limiting their utility for prediction. However, some criteria performed worse than others, in particular, we found AIC and AIC_c to generally perform poorly in our applications. We recommend the use of BIC and nMDL due to their consistent performance in our experiments (BIC and two-stage MDL were similar).

References

[1] J.-M. Agosta, J. Chandrasekar, M. Crovella, N. Taft, and D. Ting. Mixture models of endhost network traffic. In *Proc. INFOCOM Miniconference*, Turin, Italy, 2013.

[2] W. Aiello, A. Gilbert, B. Rexroad, and V. Sekar. Sparse approximations for high fidelity compression of network traffic data. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 22–35, October 2005.

[3] H. Akaike. A new look at statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723, December 1974.

[4] J. D. Brutag. Aberrant behavior detection and control in time series for network monitoring. In *Proceedings of the 14th Systems Administration Conference (LISA 2000)*, New Orleans, LA, USA, December 2000.

[5] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Info. Theory*, 52(12):5406–5425, December 2006.

[6] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *J. Sci. Computing*, 20(1):33–61, 1998.

[7] G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 2nd edition, 2006.

[9] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Trans. Networking*, 5(6):835–846, December 1997.

[10] J. P. C. L. d. Danilo Fernandes Tenório and R. T. de Sousa Jr. Greatest eigenvalue time vector approach for blind detection of malicious traffic. In *Proc. International Conf. on Forensic Computer Science (ICoFCS)*, pages 46–51, 2013.

[11] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions with matching pursuits. *Optical Engineering*, 33(7), July 1994.

[12] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[13] B. Eriksson, P. Barford, R. Bowden, M. Roughan, N. Duffield, and J. Sommers. BasisDetect : A model-based network event detection framework. In *ACM SIGCOMM Internet Measurement Conference*, Melbourne, Australia, 2010.

[14] P. D. Grünwald, I. J. Myung, and M. A. Pitt. *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*. The MIT Press, 2005.

[15] M. H. Hansen and B. Yu. *Minimum description length model selection criteria for generalized linear models*, volume 40 of *Lecture Notes–Monograph Series*, pages 145–163. Institute of Mathematical Statistics, 2003.

[16] C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

[17] A. Lakhina, M. Crovella, and C. Diot. Characterization of network-wide anomalies in traffic flows. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 201–206, 2004.

[18] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM*, pages 219–230, September 2004.

[19] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. *SIGMETRICS Perform. Eval. Rev.*, 32(1):61–72, June 2004.

[20] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Networking*, 2(1):1–15, February 1994.

[21] M. A. Little and N. S. Jones. Generalized methods and solvers for noise removal from piecewise constant signals: II. New methods. *Proc. R. Soc. A*, 471(2179), June 2011.

[22] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2nd edition, 2001.

[23] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42(5), 1989.

[24] NLANR. Abilene Trace Data. http://www.maths.adelaide.edu.au/matthew.roughan/Stuff/Abilene.tar.gz.

[25] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit : Recursive function approximation with application to wavelet decomposition. In *Asilomar Conf. on Signals, Systems and Comput.*, pages 1–5, November 1993.

[26] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Trans. Networking*, 3(3):226–244, June 1995.

[27] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Statis.*, 11(2):416–431, June 1983.

[28] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang. Experience in measuring backbone traffic variability:

Models, metrics, measurements and meaning. In *ACM SIGCOMM Internet Measurement Workshop*, 2002.

[29] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[30] A. Soule, A. Nucci, R. Cruz, E. Leonardi, and N. Taft. How to identify and estimate the largest traffic matrix elements in a dynamic environment. *SIGMETRICS Perform. Eval. Rev.*, 32(1):73–84, June 2004.

[31] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon. Providing public intradomain traffic matrices to the research community. *Computer Communication Review*, 36(1):83–86, January 2006.

[32] M. Vetterli and J. Kovačević. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[33] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan. Network anomography. In *ACM Internet Measurement Conference*, Berkeley, California, USA, October 2005.