# Measuring traffic congestion: an approach based on learning weighted inequality, spread and aggregation indices from comparison data

Gleb Beliakov[a], Marek Gagolewski[b,c], Simon James[a,1], Shannon Pace[d], Nicola Pastorello[d], Elodie Thilliez[d], Rajesh Vasa[d]

[a] *School of Information Technology, Deakin University 221 Burwood Hwy, Burwood, Victoria 3125, Australia*

[b] *Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland*

[c] *Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland*

[d] *Deakin Software and Technology Innovation Lab (DSTIL), Deakin University 221 Burwood Hwy, Burwood, Victoria 3125, Australia*

---

**Abstract**

As cities increase in size, governments and councils face the problem of designing infrastructure and approaches to traffic management that alleviate congestion. The problem of objectively measuring congestion involves taking into account not only the volume of traffic moving throughout a network, but also the inequality or spread of this traffic over major and minor intersections. For modelling such data, we investigate the use of weighted congestion indices based on various aggregation and spread functions. We formulate the weight learning problem for comparison data and use real traffic data obtained from a medium-sized Australian city to evaluate their usefulness.

*Keywords:* Aggregation functions, inequality indices, spread measures, learning weights, congestion, traffic analysis.

---

[1]Corresponding author. Tel.: +61 3 9251 7481.
*E-mail address:* sjames@deakin.edu.au.

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

## 1. Introduction

When making decisions based on sets of numerical data, as data analysts we will usually be interested in summaries that help us identify trends, central tendency, spread, etc., and which can be used to make objective comparisons. Classical operators such as the arithmetic mean and median have been recognized as special cases of much broader families of aggregation functions, which have been studied in depth in the areas of decision-making and fuzzy systems [7, 27, 35]. However, in many contexts there is also the need for a more dedicated study of summaries that indicate the variation or spread of data [23]. In particular, we have measures which evaluate the level of income inequality in economics [4, 20, 26], the evenness of species distributions in ecology [2, 32, 34], and disagreement between experts in group decision making [3, 8, 17, 1, 24].

Traffic analysis is a topic of interest across various research fields, with the core problems of understanding, modelling, predicting and reducing traffic congestion being useful not only in terms of efficient infrastructure and logistics, but also in terms of environmental impact. Traffic and network simulation has been useful in understanding different theories on flow (e.g. [29]) and also in predicting the impact of certain control measures, e.g. charging for entering central business zones [37]. Real data is either obtained from stationary sensors (cameras, vehicle detectors etc.,) [19, 39, 41] or, more recently, from GPS trajectories based on devices embedded in vehicles (especially vehicles such as taxis [36], where privacy concerns are not considered as relevant).

Decision makers and most real-time automated systems in traffic management rely on traffic volume data (e.g. see [31]) that counts the number of cars passing through an intersection over a given time interval (although across some freeway networks there may also be average speeds available [19]). Rather than volume, road users, council decision makers and traffic managers will usually be interested in the level of *congestion* experienced across a given region or large network. Whereas volume can be measured objectively, the notion of congestion is somewhat more difficult to define. It has been approached as a binary classification task in [39], i.e. where an intersection is considered to be congested when the volume exceeds a certain threshold, while in [21] an expert system was proposed that distinguishes between incidents and congestion. We are interested in developing reliable indices of congestion given over a continuous scale so that the impact of po-

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

tential improvements to the network, e.g. from road work, new highways, changes to traffic light sequences etc., can be measured. By being able to objectively measure congestion in a way that reflects the road-user experience (i.e. traffic jams and slower travel speeds), decision makers can then consider how best to reduce congestion.

Periods of high volume certainly will often correspond with drivers experiencing high levels of congestion, however in terms of the number of cars passing through an intersection, low counts can also be indicative of high congestion. While in [39] congestion prediction problem was approached as one of feature selection in the presence of correlated variables, intuitively we can recognize that the function behavior we are interested in is one whose output can tell us when, in a local area of the network, large intersections have counts below their capacity while other intersections are all busier than normal. For this we turn to inequality, spread and consensus functions, all of which provide summaries of a dataset's variation, although from slightly different perspectives.

In this contribution, we will consider weighted versions of these indices and functions toward their practical application in measuring congestion and more broadly for decision making applications. As an illustrative example of their use, we will use a subset of traffic data obtained from Brisbane City Council (in Australia) measuring the volume of traffic passing through various intersections over 5-minute intervals.

We will organize our contribution according to the following structure. In the Preliminaries section, we will give the necessary background and formulas for aggregation functions, inequality functions, spread measures, and consensus measures. In Section 3, we formulate the linear programming approaches required for fitting our simple congestion metrics to data. In Section 4, we use the traffic volume and median velocities for learning weights and measuring congestion across a subset of the network. We look at the performance of each measure in terms of Spearman correlation [33] both in fitting and for use in prediction. In Section 5, we will provide some discussion and outline some avenues for future research, before providing concluding remarks in Section 6.

## 2. Preliminaries

We consider the topics of aggregation, inequality, spread and consensus in the context of measuring traffic congestion, where inputs will usually relate

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

to a set of intersection volumes, however these are of course also relevant to multi-criteria evaluation and decision making in general.

### *2.1. Aggregation functions*

Aggregation functions [7, 27, 35], which include statistics such as the arithmetic mean and median, are useful for summarizing sets of inputs with a single value. We will be concerned primarily with non-negative inputs.

**Definition 1.** *For a given positive real interval* $[0, b]$ *and a fixed* $n \in \mathbb{N}\setminus\{1\}$*, an aggregation function* $\mathsf{A} : [0, b]^n \to [0, b]$ *is a function non-decreasing in each argument and satisfying* $\mathsf{A}(0, \ldots, 0) = 0$ *and* $\mathsf{A}(b, \ldots, b) = b$*.*

Typical examples include the arithmetic mean, i.e. the sum of inputs divided by $n$, and the median, which is the central input (or mean of the central two inputs) when the input vector is ordered in ascending or descending order. Both are examples of *averaging* aggregation functions, which for all $\mathbf{x} \in [0, b]^n$ satisfy

$$\min(\mathbf{x}) \leq \mathsf{A}(\mathbf{x}) \leq \max(\mathbf{x}).$$

In the context of congestion, averaging aggregation functions can be interpreted as measures of the volume of traffic per intersection. The property of idempotency ensures that $\mathsf{A}(a, a, \ldots, a) = a$ for all $a$, so that if the volume were the same for each intersection at a given time $t$, the output would be that volume.

Some families of aggregation functions can be defined with respect to a weighting vector so that certain inputs have a higher influence on the function output.

**Definition 2.** *A vector* $\mathbf{w} = (w_1, \ldots, w_n)$ *is called a weighting vector if* $w_j \in [0, 1]$ *for each* $j = 1, \ldots, n$ *and* $\sum_{j=1}^{n} w_j = 1$*.*

With respect to $\mathbf{w}$, we have the weighted arithmetic mean, given by

$$\mathsf{WAM}_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{n} w_j x_j.$$

The WAM allocates importance to inputs according to their source, i.e. in measuring traffic volume it may be that certain intersections are more important for influencing congestion than others. In some applications, it is useful to assign weight based on the relative size of the inputs, e.g. being able to weight busy intersections higher or lower than quiet intersections. For this we use the ordered weighted averaging operator [38], given by

$$\mathsf{OWA}_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{n} w_j x_{(j)},$$

where (.) indicates a non-increasing permutation of the inputs, $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(n)}$.

Special cases of the OWA operator, depending on the weighting vector $\mathbf{w}$ include:

***Arithmetic mean*** where all the weights are equal, i.e. all $w_j = \frac{1}{n}$;

***Maximum*** function for $\mathbf{w} = (1, 0, ..., 0)$;

***Minimum*** function for $\mathbf{w} = (0, ..., 0, 1)$;

***Median*** function for $w_j = 0$ for all $i \neq h$, $w_h = 1$ if $n = 2h - 1$ is odd, and $w_j = 0$ for all $i \neq h, h + 1$, $w_h = w_{h+1} = 0.5$ if $n = 2h$ is even.

While average volume is useful for measuring periods of high-volume, it may not correspond well with times where congestion is so high that the flow of traffic, and therefore the volume passing through each intersection, starts to decrease. We have the following concepts which relate to the distribution and variation of inputs.

### *2.2. Inequality indices*

Inequality indices have mainly been studied in the context of economic inequality [5, 18, 22] as well as having applications in ecology for measuring biodiversity and ecological evenness [11, 13, 32, 34].

A key property that characterizes inequality indices is satisfaction of the Pigou-Dalton principle [20], also referred to as the progressive transfers principle.

**Definition 3.** *For a set of inputs $\mathbf{x}$ fixed for all arguments except $x_i > x_j$, a function $f$ satisfies the Pigou-Dalton or progressive transfers principle if for all $x_i - h \geq x_{i+1} \geq \cdots \geq x_{j-1} \geq x_j + h, h > 0$ it holds that*

$$f(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n) \geq f(x_1, \ldots, x_i - h, \ldots, x_j + h, \ldots, x_n).$$

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

It ensures that if proportional input is transferred from a higher to a lower input, the output of the function will not increase. Inequality is hence at its highest when a single input is the only value above zero, and at its lowest when all inputs are equal. In an economic context, high inequality corresponds with a single individual holding 100% of the wealth, while in ecology high inequality (low evenness) indicates a single species dominating in terms of abundance.

Intuitively this is appealing for measuring traffic congestion because when traffic congestion is low, we expect a few busy intersections to experience high volume with cars freely passing through, while smaller intersections remain quiet. At times of peak congestion, smaller intersections become busier, and the traffic passing through larger intersections starts to decrease. We can even think of strategies for navigating through peak hour traffic, where we might avoid a busy intersection and use an alternative route that goes through (usually) quieter intersections if the traffic seems blocked up.

We will define inequality functions in terms of the cumulative sum [9]:

$$\mathsf{cumsum}(\mathbf{x}) = \left(x_1, x_1 + x_2, \ldots, \sum_{j=1}^{n} x_j\right),$$

and the space $\mathcal{Q}$ where

$$\mathcal{Q} = \{(q_1, \ldots, q_n) \in [0,1]^n : q_1 \geq \cdots \geq q_n, \sum_{i=1}^{n} q_i = 1\}.$$

We assume $\mathbf{x} \neq (0, 0, \ldots, 0)$ and clearly, for any $\mathbf{x} \in [0, b]^n$, it will hold that $\frac{1}{\mathsf{Sum}(\mathbf{x})}(x_{(1)}, \ldots, x_{(n)}) \in \mathcal{Q}$.

**Definition 4.** *An inequality index is a function* $\mathsf{I} : \mathcal{Q} \to [0,1]$ *such that: for all* $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$, *if* $\mathsf{cumsum}(\mathbf{q}) \leq \mathsf{cumsum}(\mathbf{q}')$, *then* $\mathsf{I}(\mathbf{q}) \leq \mathsf{I}(\mathbf{q}')$, *with* $\inf_{\mathbf{q}\in\mathcal{Q}} \mathsf{I}(\mathbf{q}) = 0$ *and* $\sup_{\mathbf{q}\in\mathcal{Q}} \mathsf{I}(\mathbf{q}) = 1$.

The supremum for an inequality index is obtained for $(1, 0, 0, \ldots, 0)$ while the infimum occurs at $(1/n, \ldots, 1/n)$. Note that each $I$ satisfies the Pigou-Dalton principle. Of course, indices are often defined on spaces other than $\mathcal{Q}$ and in such cases may result in outputs over a different range.

One of the most prevalent inequality indices used is the Gini index [26],

$$\mathsf{G}(\mathbf{x}) = \frac{1}{2n^2 \cdot \mathsf{AM}(\mathbf{x})} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|,$$

Over $\mathcal{Q}$ it is expressed [40]:

$$\mathsf{G}(\mathbf{q}) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}|q_i - q_j|.$$

An interesting family of inequality indices was proposed in [5] in terms of a dual composition of welfare functions. A welfare function is an aggregation function $\mathsf{W} : [0, b]^n \to [0, b]$ that satisfies Schur-concavity, which means that the function increases if input is proportionally transferred from higher to lower inputs. This property can be ensured by using $\mathsf{OWA}$ functions with respect to weighting vectors satisfying $0 < w_1 < w_2 < \cdots < w_n$.

Let $\mathsf{W}^d_{\mathbf{w}}$ denote the dual of the welfare function $\mathsf{W}_{\mathbf{w}}$, i.e. the $\mathsf{OWA}$ with the same weights in reverse order. Then, the following will be an inequality function.

$$\mathsf{I}_{\mathbf{w}}(\mathbf{q}) = \frac{\mathsf{W}^d_{\mathbf{w}}(\mathbf{q}) - \mathsf{W}_{\mathbf{w}}(\mathbf{q})}{2}. \tag{1}$$

The Gini index $\mathsf{G}(\mathbf{x})$ corresponds with weights $w_i = \frac{2n+1-2i}{n^2}$.

We also make mention of the Simpson domination index [32], which is given by

$$\mathsf{S}(\mathbf{q}) = \sum_{j=1}^{n} q_j^2,$$

where $q_i$ is the proportional held by the $i$-th input. It represents inequality in terms of dominance, however we note that it can also be seen as a variation on the weighted arithmetic mean, where each $w_j$ is equal to $q_j$.

As well as inequality functions being intuitively appealing, we may also find it useful to consider spread measures, a broader class of functions based on input differences, for measuring congestion.

### *2.3. Spread measures*

The concept of spread measures and their relationship to aggregation functions was formalized in [23]. We first denote the vector of iterated differences

$$\mathsf{diff}(\mathbf{x}) = (x_{(1)} - x_{(2)}, x_{(2)} - x_{(3)}. \ldots, x_{(n-1)} - x_{(n)}),$$

Note that $\mathsf{diff}(\mathbf{x})$ is a vector of $n-1$ components.

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

**Definition 5.** *An (absolute) symmetric spread measure is a function* $\mathsf{V} : [0,b]^n \to [0,\infty]$ *such that for each* $\mathbf{x}, \mathbf{x}'$ *if it holds that* $\mathsf{diff}(\mathbf{x}) \leq \mathsf{diff}(\mathbf{x}')$*, then* $\mathsf{V}(\mathbf{x}) \leq \mathsf{V}(\mathbf{x}')$*, and which satisfies* $\inf_{\mathbf{x}\in[0,b]^n} \mathsf{V}(\mathbf{x}) = 0$.

The latter condition is equivalent to stating that $\mathsf{V}(a, a, \ldots, a) = 0$ for all $a \in [0, b]$.

Spread measures, which include the statistical standard deviation and range, can hence be constructed from aggregation functions on the vector of iterated differences $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{n-1}) = \mathsf{diff}(\mathbf{x})$.

### *2.4. Consensus measures*

In the context of decision making, consensus measures are usually defined as functions satisfying unanimity (all inputs the same results in the highest output), anonymity and neutrality (symmetry with respect to the experts and alternatives) [16, 25]. Another property that is often desired and distinguishes consensus measures from other types of variation-based functions is that of *maximum dissension*, i.e. that when inputs are partitioned into two groups at extreme ends of the scale, the function meets its minimum. We can compare this to inequality, where the partition of extreme evaluations is not equal. Associated with this property is monotonicity with respect to the majority [8], where consensus should increase if the distance between a majority of inputs and another is reduced. We will consider the negation of one such function, so that it instead models 'disagreement' and can be interpreted consistently with inequality and spread. It is based on the distance between each input and the median.

$$\mathsf{Cons}(\mathbf{x}) = \sum_{i=1}^{n} w_i \left| x_i - \mathsf{Med}(\mathbf{x}) \right|.$$

Such a function is also evidently a spread measure, since it will be shift-invariant and increasing with respect to increases to the iterated differences.

## 3. Learning spread measure and inequality weights from comparison data

To use weighted functions for assessing the level of congestion, we need a way of choosing the appropriate weights. Functions such as the Gini index are formulated with respect to fixed (equal) weights, however in our context it is likely the case that some intersections will have a higher influence on the level of congestion, e.g. due to the topology of the road network.

We assume datasets consisting of $m \times n$ matrices, where each row $\mathbf{x}_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,n})$ denotes the number of cars passing through each of the $n$ intersections over the $k$-th time interval. For weighted aggregation functions, along with the observed inputs $\mathbf{x}_k$, ideally we also have observed outputs $y_k, k = 1, \ldots, m$ and we can determine weights by minimization of the least absolute deviation (LAD) of residuals [6, 15].

For a function $f_{\mathbf{w}}$ dependent on $\mathbf{w}$ and with $W$ the set of potential weighting vectors, we have

$$\underset{\mathbf{w} \in W}{\text{Minimize}} \quad \sum_{k=1}^{m} \left| f_{\mathbf{w}}(\mathbf{x}_k) - y_k \right|, \tag{2}$$

subject to any desired constraints.

In our context, congestion cannot be observed directly, and so we instead have to rely on the estimated median velocity over the network at each time step. Although it makes sense for the median velocity to be correlated negatively with congestion, it might not be the case that changes in velocity are commensurate with changes in our approximation $f_{\mathbf{w}}(\mathbf{x}_k)$ and so we consider approaches based on finding functions that attempt to satisfy $f_{\mathbf{w}}(\mathbf{x}) \geq f_{\mathbf{w}}(\mathbf{x}')$ whenever the velocity corresponding with $\mathbf{x}$ is lower than that at $\mathbf{x}'$ (see e.g. [10, 12]). In other words, we aim to define functions that maintain the relative ordering according to observed pairs of outputs.

For each $\mathbf{x}_k$, we denote the median velocity over the entire network by $y_k$ and construct a set of comparisons $\mathcal{P} \subseteq \{(a, b) : a, b \in \{1, \ldots, m\}\}$ such that $(a, b) \in \mathcal{P}$ indicates that we desire $f_{\mathbf{w}}(\mathbf{x}_a) > f_{\mathbf{w}}(\mathbf{x}_b)$. For some indices, this will correspond with $y_a > y_b$, e.g. when we have low inequality we expect the median velocity to be lower, whereas for other indices it will be when $y_a < y_b$, e.g. when the volume increases we expect the median velocity to decrease. We use median velocity since it is the best indicator of congestion available to us, however we note that the approaches set out below could similarly be adopted if we employed some other measure of congestion, e.g. observations from traffic cameras where experts assess the congestion as being low, medium, very high etc.

### *3.1. Weighted congestion indices based on aggregation*

We will formulate the required constraints for approximating a weighted arithmetic mean from comparison data. Since a weighted mean of traffic volumes can be interpreted as volume per intersection, we expect that increases to our WAM should correspond with decreases in velocity, i.e. for each $y_a < y_b$, we want it to hold that $\mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_a) > \mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_b)$. We hence include the constraint

$$w_1(x_{a,1} - x_{b,1}) + w_2(x_{a,2} - x_{b,2}) + \cdots + w_n(x_{a,n} - x_{b,n}) - r^+_{a,b} + r^-_{a,b} = \varepsilon$$

where $\varepsilon$ is the minimum difference required, $r^+_{a,b}$ indicates the extent to which $\mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_a)$ is greater than $\mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_b)$ and $r^-_{a,b}$ is how much the desired constraint is violated. The $\varepsilon$ sets a minimum value upon which $\mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_a)$ is deemed greater (so that we do not obtain a function that favors giving a constant output). In our objective function, we then aim to minimize the $r^-_{a,b}$ and maximize the $r^+_{a,b}$ to get as much separation as possible. We construct a set of comparisons $\mathcal{P}$ such that $(a, b) \in \mathcal{P}$ indicates $y_a < y_b$. Depending on the dataset, in practice we can use the set of all $m(m-1)/2$ comparisons, or construct a smaller set to reduce the number of constraints. In the case of congestion, we would not necessarily assume that a low speed at 9:30 in the morning would be associated with the same traffic features that cause a low speed at 18:30 in the evening, and so we instead focus on changes in velocity between consecutive intervals, i.e. if the the median velocity decreases, then we predict the average volume to have increased, while if the velocity increases we predict the volume to have decreased.

Along with conditions that ensure $\mathbf{w}$ is a weighting vector, we hence obtain the following.

$$\begin{aligned}
\underset{\mathbf{w}\in W}{\text{Maximize}} \quad & \sum_{(a,b)\in\mathcal{P}} r^+_{a,b} - \lambda r^-_{a,b}, \\
\text{s.t.} \quad & \mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_a) - \mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_b) - r^+_{a,b} + r^-_{a,b} = \varepsilon, \forall (a,b) \in \mathcal{P}, \\
& \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0, j = 1, \ldots, n \\
& r^+_{a,b}, r^-_{a,b} \geq 0, \forall (a,b) \in \mathcal{P}. \qquad (3)
\end{aligned}$$

The value $\lambda$ is a penalty parameter and can be used to control how much we allow violated pairwise judgements to be compensated for by high satisfaction (i.e. high values of $r^+_{a,b}$) of others. Removing $r^+_{a,b}$ altogether (or making

$\lambda$ very high) would mean we focus only on limiting the violations whereas here we attempt to make $|\mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_a) - \mathsf{WAM}_{\mathbf{w}}(\mathbf{x}_b)|$ as large as possible if $(a, b) \in \mathcal{P}$.

We can similarly formulate the problem for congestion evaluations based on OWA functions. We need only to reorder each of the input vectors so that they are in decreasing order.

### *3.2. Weighted congestion indices based on inequality*

Weights for the inequality measures of the form given by Eq. (1) can be learned from data using essentially the same method as for aggregation functions. However, rather than fitting to the terms $x_{k,j}$, we fit to $u_{k,j}$ where

$$u_{k,j} = x_{k,(j)} - x_{k,(n-j+1)},$$

and hence we will have the constraints of the form

$$w_1(u_{a,1} - u_{b,1}) + w_2(u_{a,2} - u_{b,2}) + \cdots + w_n(u_{a,n} - x_{b,n}) - r^+_{a,b} + r^-_{a,b} = \varepsilon$$

for each $(a, b) \in \mathcal{P}$. We remind that (.) in the index of $x_{k,(j)}$ denotes the inputs being arranged in descending order for each time index in the calculation of $u_{k,j}$.

The weights applied to each term $u_{k,j}$ will then correspond with the $w_j$ associated with the dual of the welfare function. Since welfare functions require increasing weights, the dual weights are required to be decreasing, and so we need to add constraints,

$$w_j > w_{j+1}, \qquad j = 1, \ldots, n-1.$$

In the case of inequality indices based on welfare functions, we assume that decreases in median velocity correspond with decreases in median velocity, i.e. that congestion increases when the inequality is reduced. For each $y_a > y_b$ we therefore will have $(a, b) \in \mathcal{P}$.

### *3.3. Weighted congestion indices based on spread measures*

Spread measures will have the same expected relationship between inputs and outputs, with decreases to velocity also decreasing the spread. We will consider spread measures constructed from aggregations of the iterated differences vector and from the set of $n(n-1)/2$ absolute differences as observed in functions like the Gini index.

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

In the former case, we have a vector of $(n-1)$ arguments and we replace $x_{k,j}$ terms with $\delta_{k,j}$ in the LP formulation. We can consider $\mathsf{WAM}_{\mathbf{v}}(\boldsymbol{\delta})$ where $\boldsymbol{\delta} = \mathsf{diff}(\mathbf{x})$ (with the inputs ordered decreasingly before calculating the differences), $\mathbf{v} = (v_1, v_2, \ldots, v_{n-1})$, then for each $y_a > y_b$ we will have $(a, b) \in \mathcal{P}$ and add the constraint

$$v_1(\delta_{a,1} - \delta_{b,1}) + v_2(\delta_{a,2} - \delta_{b,2}) + \cdots + v_{n-1}(\delta_{a,n-1} - \delta_{b,n-1}) - r^+_{a,b} + r^-_{a,b} = \varepsilon.$$

In this case there will be no special conditions on the weights, although if we want our spread measures to strictly increase for increasing $\boldsymbol{\delta}$ we can add constraints ensuring $v_j > 0$. In interpreting $\mathbf{v}$, high weights allocated to lower $j$ indicate that the function is more affected by changes to the high inputs while high weights for higher $j$ indicate that the function changes more with respect to low inputs.

As a different kind of spread measure, we can consider input vectors consisting of all absolute differences $|x_{k,i} - x_{k,j}|$, and hence find weights $W_{i,j}$ such that our congestion is measured by

$$\sum_{i=1}^{n} \sum_{j=1}^{n} W_{i,j} |x_i - x_j|.$$

The weight fitting procedure here is no different than that applied to the case of $\mathbf{v}$, however we will be fitting to $n(n-1)/2$ weights rather than $(n-1)$. It can be considered as a weighted version of the Gini index, with high weights being indicative of either one or both of the corresponding intersections being important in influencing congestion.

With considerably more weights, the function will be more flexible and this then comes with the potential to overfit the data, so this needs to be taken into account when analysing the weighting vectors that result from the fitting process.

### *3.4. Weighted congestion indices based on disagreement*

Lastly we will consider learning weighted functions of disagreement (based on consensus measures). Once again, we expect the disagreement to be high when congestion is low, since this represents the case of half of the intersections being busy and half having fewer cars passing through. In this case, we transform the inputs by calculating their distance to the median output, so that we have

$$\eta_{k,j} = |x_{k,j} - \mathsf{Med}(\mathbf{x}_k)|.$$

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

We then fit to $\mathbf{w}$, so that for each $y_a > y_b$ we have $(a, b) \in \mathcal{P}$ and the constraint

$$w_1(\eta_{a,1} - \eta_{b,1}) + w_2(\eta_{a,2} - \eta_{b,2}) + \cdots + w_n(\eta_{a,n} - \eta_{b,n}) - r^+_{a,b} + r^-_{a,b} = \varepsilon.$$

The weights are then interpreted with respect to the $j$-th input, with high weights indicating that changes to $x_j$ affect the output more.

## 4. Evaluation of various metrics on the Brisbane traffic dataset

We obtained volumetric data (number of cars passing through an intersection over a 5-minute interval observed using vehicle detecting road sensors) and median velocity data (extracted from blue-tooth GPS data for a sample of 3000 cars over the entire network) from the Brisbane City Council for 8 weekdays from September 5 to September 14[2], 2016. We are interested in illustrating how our methods could be applied, and so we consider 5 randomly chosen geographical regions over which we want measure the local congestion and apply our methods.

The intersections selected are shown in Fig. 1(b), grouped into regions (chosen arbitrarily and based on rectangular bounds in terms of GPS coordinates) consisting of 33, 60, 42, 51 and 43 intersections respectively with varying density and proximity to the city center.

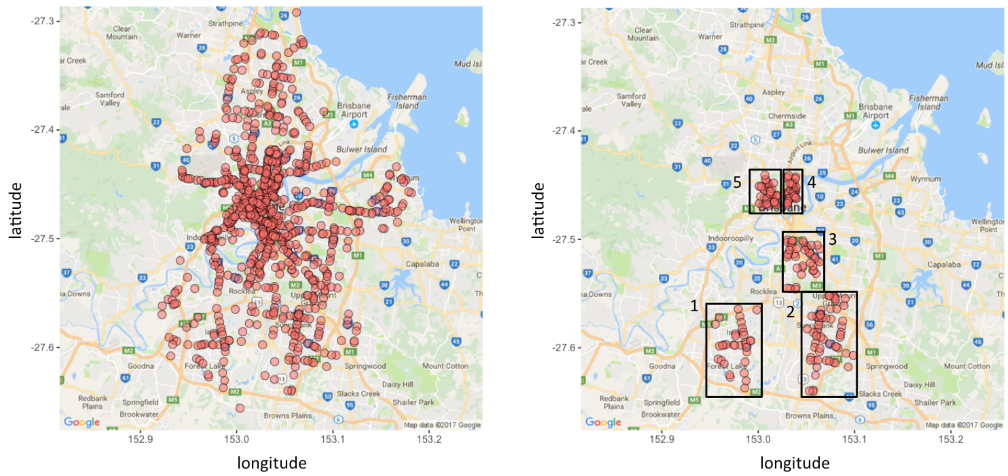
The median velocity we used as a proxy for congestion was obtained based on travel times and distances for the same period. It is worth noting that this gives the estimated median velocity over the entire network, and so the smaller regions we have may or may not have volumes that are representative of the entire network and consistent with the changes observed in the velocity data[3].

Fig. 2(a) gives an example of the velocity time series data in km/h over a whole day. We can see high variability, particularly at night time (each time index indicates a 5 minute period), due in part to the low number of cars in the network over these times. For our experiments we smoothed the

[2]We restricted our selection to these particular days due to reliability of the velocity data and to avoid using weekends and school holidays, which are likely to have less extreme congestion peaks.

[3]All code used for fitting experiments and individual results are available from http://aggregationfunctions.wordpress.com, while datasets used can be obtained by contacting the authors. All techniques were implemented in R [30].

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919. doi:10.1016/j.asoc.2017.07.014



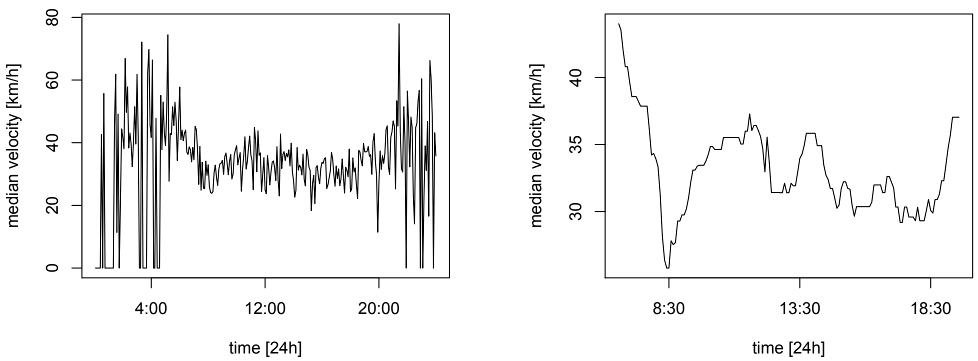

(a) all intersections (b) intersections in selected regions

Figure 1: Roadmap of Brisbane city with all intersections collecting traffic volume data (a) and the five regions we used for our fitting experiments (b). In the five regions there are 33, 60, 42, 51 and 43 intersections respectively. Maps were created using ggmap [28] in R.

data by taking the median of the previous 12 time steps and using only the 157 data points between 6:30 and 19:30 (shown in Fig. 2(b)). We used this time window because we are mainly interested in observing the periods of high congestion. We opt for the median over the mean to avoid influence of outliers and use the previous 12 time-steps rather than basing our calculation at the center so that our functions could theoretically be used at each step (in real time) when evaluating congestion for new unobserved instances. Using 12 time-steps does mean that if a drastic change in congestion occurred, it might not be noticed until up to half an hour has passed, however using 'smoother' data is preferable for us since we are not looking to model the median velocity curve exactly, and furthermore for our comparison-based fitting technique we desire differences between time-steps to be reliable.

We applied the same smoothing technique to the traffic volume count data for each of the intersections and used the same time window. We then applied the fitting techniques for the weighted arithmetic mean (WAM), weighted inequality indices based on welfare functions ($\mathsf{I}_\mathbf{w}$), weighted spread measures ($\mathsf{V}$), and weighted disagreement functions based on consensus measures (Cons) to each of the following:

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919. doi:10.1016/j.asoc.2017.07.014




(a) unsmoothed over entire day

(b) smoothed over 6:30 – 19:30 time window

Figure 2: Unsmoothed data over a whole day (a) and data smoothed using the median of the previous 12 time-steps over time indices 78 to 234 (corresponding with 6:30am to 7:30pm) shown in (b).

1. Inputs $x_{k,j}$ indicating the (smoothed) number of cars passing through intersection $j$ at the $k$-th time index;
2. Proportional inputs $\mathbf{q}$ obtained by using $q_{k,j} = x_{k,j}/\sum_{i=1}^{n} x_{k,i}$. Each input then represents the proportion of cars in a given region passing through that intersection;
3. Relative inputs scaled to $[0, 1]$ by dividing through by the largest observed $x_{k,j}$ for each $j$, i.e. so that each input represents the percentage of maximum capacity the intersection is experiencing in terms of volume;
4. Proportional values obtained from the relative inputs.

In order to obtain the set of comparisons $\mathcal{P}$, we used the data in sequence and considered only each pair of neighboring time-steps. Each fitting exercise hence used 156 constraints in addition to any required by the weighting vectors. We used the value of $\lambda = 5$ to strike a balance between over-enforcing the correspondence between changes in median velocity and increases or decreases in the function, and allowing for violations to be compensated for by higher differences at other times). We set the value of $\varepsilon$ to 0.00001, which we also set as the numerical threshold when requiring weight inequalities $w_j > w_{j+1}$ to be satisfied. We did not use the weighted Gini-based spread measures since the number of parameters would be in excess of 10 times the number of training instances and hence not likely to give reliable results.

To evaluate the performance of each of our weighted congestion metrics, we use the Spearman correlation coefficient $\rho$ with the smoothed median velocity since it gives us, intuitively, the degree to which one variable can be expressed as a strictly increasing (positive coefficient) or decreasing (negative coefficient) transform of the other. It is worth noting that the fitting approach does not necessarily optimize for this value due to the ability of high differences for some pairs to compensate for violations of the ordering elsewhere.

### *4.1. Average fitting performance over all days*

For the following tables and results, we recall the following abbreviations and notation:

| | |
|---|---|
| reg | The region as indicated in Fig.1(b). |
| WAM | Weighted traffic volume |
| $\mathsf{I}_\mathbf{w}$ | Weighted inequality index |
| V | Weighted spread |
| Cons | Weighted disagreement |
| volume | Raw volume |

Columns marked **q** indicate the same index applied to proportional values.

We note that as all methods are linear (in fact the the same fitting function is used, only with different transformations of the data based on the desired type of index), the implementation is quite fast and scales with the number of sensors and constraints. For example, for the weighted mean on region 1 (33 sensors) the fitting for one day (157 instances) takes approximately a tenth of a second when run on a 2.7GHz processor with 8 GB of RAM.

Table 1 gives the average performance for each of the functions and input-type combinations over the different regions we used. Each value represents the average Spearman correlation over the 8 days, i.e. the arithmetic mean of the $\rho$ statistic calculated each day. For a day with 157 time steps considered as is the case here, Spearman values will be statistically significant (i.e. the correlation is different from 0) at a 0.05 significance level whenever $|\rho|$ is above approximately 0.16, so these averages can be interpreted with this in mind[4].

For WAM, this value is expected to be negative, since we anticipate congestion to increase with volume and for this to reduce the median speed experienced across the network. This was the only function fit with this assumption, i.e. in determining the order of $(a, b)$ when including the comparison pair in $\mathcal{P}$. On the other hand, note that for the weighted arithmetic mean of proportional inputs $\mathsf{WAM}(\mathbf{q})$, we anticipated the function to behave as a re-weighted version of the Simpsons's index and hence expect a positive correlation with velocity.

Table 1: Average $\rho$ over 8 days for functions fit using median velocity comparison data.

Standard count data

| reg | WAM | $(\mathbf{q})$ | $\mathsf{I}_{\mathbf{w}}$ | $(\mathbf{q})$ | V | $(\mathbf{q})$ | Cons | $(\mathbf{q})$ | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.748 | 0.412 | -0.590 | 0.410 | -0.281 | 0.458 | -0.140 | 0.482 | -0.731 |
| 2 | -0.749 | 0.701 | -0.693 | 0.680 | -0.170 | 0.598 | 0.186 | 0.655 | -0.710 |
| 3 | -0.728 | 0.689 | -0.690 | 0.713 | 0.271 | 0.668 | 0.013 | 0.726 | -0.763 |
| 4 | -0.760 | 0.592 | -0.662 | 0.320 | 0.198 | 0.631 | 0.257 | 0.709 | -0.738 |
| 5 | -0.697 | 0.640 | -0.493 | 0.630 | -0.056 | 0.599 | 0.011 | 0.620 | -0.658 |
| avg | -0.736 | 0.607 | -0.626 | 0.550 | -0.008 | 0.591 | 0.066 | 0.638 | -0.720 |

Counts transformed relative to highest volume over each day

| reg | WAM | $(\mathbf{q})$ | $\mathsf{I}_{\mathbf{w}}$ | $(\mathbf{q})$ | V | $(\mathbf{q})$ | Cons | $(\mathbf{q})$ | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **-0.751** | 0.452 | 0.282 | 0.601 | 0.441 | 0.670 | 0.356 | 0.613 | -0.731 |
| 2 | **-0.758** | 0.691 | 0.476 | 0.644 | 0.571 | 0.672 | 0.488 | 0.655 | -0.710 |
| 3 | -0.735 | 0.699 | 0.669 | 0.742 | 0.668 | 0.731 | 0.729 | **0.770** | -0.763 |
| 4 | **-0.769** | 0.686 | 0.501 | 0.664 | 0.620 | 0.702 | 0.697 | 0.751 | -0.738 |
| 5 | **-0.724** | 0.674 | 0.357 | 0.691 | 0.550 | 0.678 | 0.570 | 0.700 | -0.658 |
| avg | **-0.747** | 0.640 | 0.457 | 0.668 | 0.570 | 0.690 | 0.568 | 0.698 | -0.720 |

* Bold denotes best performing for each region (across both standard and relative-transformed inputs).

While some reasonably high correlations are observed for each of the spread indices, overall the best results were obtained for raw volume and weighted volume ($\mathsf{WAM}(\mathbf{x})$). In some cases with the raw data, a negative

[4]Statistical significance for Spearman correlation values can be calculated using cor.test() in `R` and is based on the algorithm in [14].

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

correlation was observed for some of the congestion indices, which is not overly surprising since in such cases high differences for the transformed input would occur at times where the volume is highest. To some degree, such of these indices hence incorporate both volume and spread, although the fitting procedure is optimized toward increases to velocity corresponding with increases in spread.

We can suggest that it is neither volume nor spread alone that may characterize congestion, but rather some combination of the two. Very low volume, experienced at every intersection, will obviously correspond with low congestion but also with high equality amongst the inputs. On the other hand, it is likely that even when the data is high in terms of inequality, if the volume is very high, it is still possible that it may indicate a period of high congestion.

Table 2 observes the effect of providing an 'adjusted' volume based on the level of congestion in a way that the output is still high if volume is high, but that it is decreased if the spread or inequality is high. For a given inequality or spread function $\mathsf{F}$, we divide volume by $(1+0.5\mathsf{F})$. This adjusted volume is not optimized for, and so of course there would be alternative constructions that provide better results. We hence include these results only to demonstrate the potential for such a combined index.

For $\mathsf{WAM}(\mathbf{x})$, the index was calculated by dividing volume by $(1-0.5\mathsf{F})$ while for the rest volume was divided by $(1+0.5\mathsf{F})$ where $\mathsf{F}$ is the fitted index presumed to positively correlate with median velocity. We see that in most cases this improved the value of Spearman's $\rho$. Best results were usually obtained where the function $\mathsf{F}$ originally had achieved reasonable results in fitting or prediction by itself. Fig. 3 provides a visual of this adjusted index, with depictions of the time-series for median velocity, volume, a fitted spread index and the adjusted measure for Region 3 on the 12th of September.

A negative correlation exists between velocity (a) and volume (b), which in this case is $\rho = -0.698$. The correlation between spread (calculated from proportional values of relative volume) and the median velocity is positive with $\rho = 0.712$ and when the volume is adjusted by spread the Spearman correlation becomes $-0.758$.

### *4.2. Using weighted functions for predicting congestion*

We now turn to the question of whether these indices can be used beyond the training data for prediction. Table 3 gives the average performance of each of the functions and input-type combinations when using the same day of the previous week for learning the weighting vector and then using this for

Table 2: Average Spearman correlation over 8 days for functions fit using median velocity comparison data using an adjusted volume index.

Standard count data

| reg | WAM | $(\mathbf{q})$ | $\mathsf{I}_\mathbf{w}$ | $(\mathbf{q})$ | V | $(\mathbf{q})$ | Cons | $(\mathbf{q})$ | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.747 | -0.743 | -0.727 | *-0.757* | *-0.756* | **-0.771** | -0.729 | *-0.757* | -0.731 |
| 2 | -0.730 | -0.746 | -0.720 | -0.712 | -0.745 | -0.751 | -0.694 | -0.739 | -0.710 |
| 3 | -0.754 | **-0.788** | -0.734 | -0.766 | -0.744 | -0.769 | -0.715 | -0.764 | -0.763 |
| 4 | -0.747 | -0.747 | -0.544 | -0.734 | -0.754 | **-0.784** | -0.711 | -0.765 | -0.738 |
| 5 | -0.677 | -0.712 | -0.645 | -0.698 | -0.644 | -0.687 | -0.703 | -0.708 | -0.658 |
| avg | -0.731 | *-0.747* | -0.674 | -0.733 | -0.729 | **-0.753** | -0.711 | *-0.747* | -0.720 |

Counts transformed relative to highest volume over each day

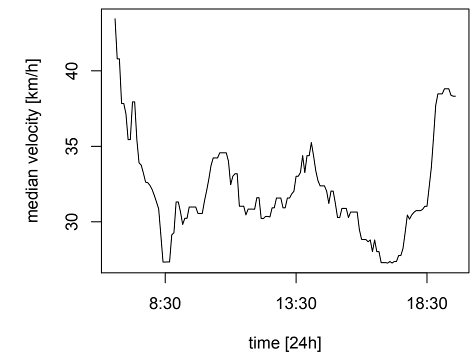
| reg | WAM | $(\mathbf{q})$ | $\mathsf{I}_\mathbf{w}$ | $(\mathbf{q})$ | V | $(\mathbf{q})$ | Cons | $(\mathbf{q})$ | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.749 | -0.742 | -0.705 | -0.722 | -0.727 | -0.739 | -0.734 | -0.742 | -0.731 |
| 2 | -0.740 | -0.749 | -0.686 | -0.699 | -0.702 | -0.702 | -0.707 | -0.715 | -0.710 |
| 3 | -0.757 | *-0.785* | -0.762 | -0.767 | -0.769 | *-0.776* | *-0.776* | *-0.774* | -0.763 |
| 4 | -0.750 | -0.757 | -0.714 | -0.730 | -0.746 | -0.748 | -0.765 | -0.760 | -0.738 |
| 5 | -0.690 | -0.719 | -0.722 | -0.704 | -0.705 | -0.695 | **-0.725** | -0.709 | -0.658 |
| avg | -0.737 | *-0.750* | -0.718 | -0.724 | -0.730 | -0.732 | -0.741 | -0.740 | -0.720 |

* Bold denotes best performing for each region (if better than previously obtained best result in Table 1) and italics indicates improvement (or equal) over best result in Table 1.

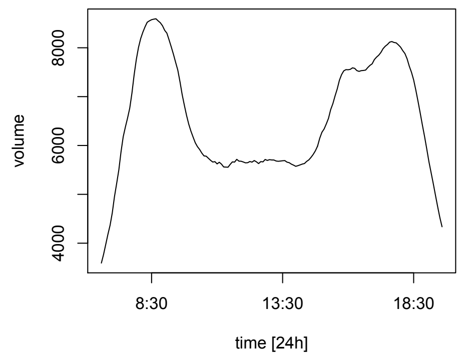
predicting the congestion based on the intersection traffic volumes at each time-step. Once again, all functions except $\mathsf{WAM}(\mathbf{x})$ were fit assuming a positive correlation with median velocity.

Overall, volume and weighted volume produce the best results, however we once again can look at the performance of adjusted volume indices, which we provide in Table 4.

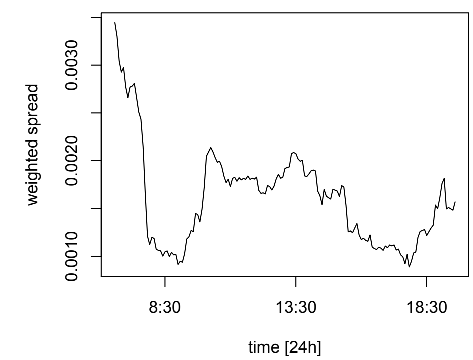
We again observe improvements in almost all cases, with the best results on average being achieved using an inequality index based on welfare functions of relative volume. This is interesting given its lower absolute correlation when used alone, although perhaps this is because it captures aspect of the data that are less correlated with volume and therefore is less redundant when used with the volume index.



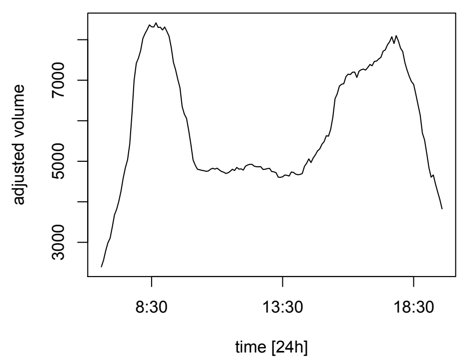

(a) velocity




(b) volume, $\rho = -0.698$




(c) weighted spread, $\rho = 0.712$




(d) adjusted volume $\rho = -0.758$

Figure 3: Comparison of smoothed time series data for velocity, volume, weighted spread (using proportions calculated from relative volume) and an adjusted volume index.

### *4.3. Using weighted spread measures over aggregated regions*

From the performance of using volume alone, we can observe that the traffic experienced in Region 5 appears to have less influence on the median velocity experienced throughout the entire network, while Region 3 is perhaps a key area influencing congestion. It is then possible that poorer performance of some of the congestion indices could be due to one or other of the regions not being highly correlated with the congestion experienced across the entire system.

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

Table 3: Average Spearman correlation over 3 days when predicting congestion from the same day the previous week.

Standard count data

| reg | WAM | $(\mathbf{q})$ | $\mathsf{I}_{\mathbf{w}}$ | $(\mathbf{q})$ | V | $(\mathbf{q})$ | Cons | $(\mathbf{q})$ | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.704 | 0.294 | -0.539 | 0.486 | -0.376 | 0.313 | -0.250 | 0.408 | -0.707 |
| 2 | -0.678 | 0.497 | -0.655 | 0.661 | -0.358 | 0.437 | 0.020 | 0.617 | -0.666 |
| 3 | -0.693 | 0.734 | -0.625 | 0.719 | 0.004 | 0.533 | -0.021 | 0.740 | -0.718 |
| 4 | -0.648 | 0.470 | -0.609 | 0.221 | -0.059 | 0.387 | 0.035 | 0.497 | **-0.662** |
| 5 | -0.592 | 0.578 | -0.438 | 0.567 | -0.210 | 0.388 | -0.073 | 0.550 | -0.579 |
| avg | -0.663 | 0.515 | -0.573 | 0.531 | -0.200 | 0.411 | -0.058 | 0.562 | -0.666 |

Counts transformed relative to highest volume over each day

| reg | WAM | $(\mathbf{q})$ | $\mathsf{I}_{\mathbf{w}}$ | $(\mathbf{q})$ | V | $(\mathbf{q})$ | Cons | $(\mathbf{q})$ | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **-0.711** | 0.378 | 0.296 | 0.697 | 0.512 | 0.659 | -0.100 | 0.568 | -0.707 |
| 2 | **-0.693** | 0.523 | 0.501 | 0.636 | 0.527 | 0.623 | 0.419 | 0.615 | -0.666 |
| 3 | -0.705 | 0.726 | 0.473 | **0.744** | 0.569 | 0.673 | 0.182 | 0.642 | -0.718 |
| 4 | -0.645 | 0.537 | 0.431 | 0.597 | 0.528 | 0.564 | 0.572 | 0.652 | **-0.662** |
| 5 | -0.612 | 0.567 | 0.247 | **0.692** | 0.542 | 0.620 | 0.217 | 0.519 | -0.579 |
| avg | **-0.673** | 0.546 | 0.389 | **0.673** | 0.536 | 0.628 | 0.258 | 0.599 | -0.666 |

* Bold denotes best performing for each region (across both standard and relative-transformed inputs).

We now investigate whether congestion can still be evaluated by treating each region as if it were a large intersection, i.e. by considering the total volume across each of the regions. In this sense, we are less likely to have 'major' and 'minor' intersections and of course fewer variables and less flexible functions, however at this level we may be able to capture behavior that is more consistent with the (non-localized) velocity data.

Table 5 gives the results for predicting the congestion based on the same day of the previous week, averaged over the three days. For these experiments, since we deal with fewer variables, we also take into account the weighted Gini index, as it now becomes a function of just 10 variables.

The first row gives the value of $\rho$ for the training data (from the same day of the previous week), the second gives the value for the test data, and the third provides the output for the adjusted index (using the same method of combining velocity with the congestion indices).

We observe similar trends here as with predicting based on each region,

Please cite this paper as: Beliakov G., Gagolewski M., James S., Pace S., Pastorello N., Thilliez E., Vasa R., Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data, *Applied Soft Computing* **67**, 2018, pp. 910-919.
doi:10.1016/j.asoc.2017.07.014

Table 4: Average Spearman correlation over 3 days when predicting congestion from the same day the previous week using an adjusted volume index.

Standard count data

| reg | WAM | (**q**) | $\mathsf{I}_\mathbf{w}$ | (**q**) | V | (**q**) | Cons | (**q**) | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.704 | -0.672 | *-0.723* | *-0.736* | -0.652 | **-0.737** | -0.651 | -0.705 | -0.707 |
| 2 | -0.669 | -0.676 | -0.684 | -0.668 | -0.646 | -0.654 | -0.644 | **-0.696** | -0.666 |
| 3 | -0.706 | *-0.752* | *-0.750* | *-0.746* | -0.667 | -0.706 | -0.726 | -0.743 | -0.718 |
| 4 | -0.650 | -0.659 | -0.451 | -0.660 | -0.556 | -0.631 | -0.592 | -0.643 | -0.662 |
| 5 | -0.578 | -0.604 | -0.616 | -0.605 | -0.544 | -0.555 | -0.652 | -0.597 | -0.579 |
| avg | -0.662 | *-0.673* | -0.645 | *-0.683* | -0.613 | -0.657 | -0.653 | *-0.677* | -0.666 |

Counts transformed relative to highest volume over each day

| reg | WAM | (**q**) | $\mathsf{I}_\mathbf{w}$ | (**q**) | V | (**q**) | Cons | (**q**) | volume |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.706 | -0.683 | *-0.713* | *-0.730* | **-0.737** | *-0.721* | -0.707 | *-0.713* | -0.707 |
| 2 | -0.678 | -0.661 | -0.669 | -0.670 | -0.664 | -0.658 | -0.666 | -0.663 | -0.666 |
| 3 | -0.716 | *-0.760* | **-0.766** | *-0.746* | -0.719 | -0.732 | *-0.744* | *-0.759* | -0.718 |
| 4 | -0.647 | -0.634 | -0.652 | -0.660 | **-0.690** | *-0.681* | *-0.667* | *-0.662* | -0.662 |
| 5 | -0.587 | -0.620 | -0.677 | -0.639 | -0.622 | -0.624 | -0.666 | -0.638 | -0.579 |
| avg | -0.667 | -0.672 | **-0.695** | *-0.689* | *-0.686* | *-0.683* | *-0.690* | *-0.687* | -0.666 |

* Bold denotes best performing for each region (if better than previously obtained best result in Table 1) and italics indicates improvement (or equal) over best result in Table 1.

with the mean performing best on its own while the remaining indices can be used in conjunction with volume to boost the fitting correlation. We also note improved correlations with the median velocity in combining the volumes for each region.

## 5. Discussion and future work

Here we have proposed methods for learning weighted spread and inequality indices and validated their potential using a small real-world dataset. In the process, a number of potential improvements that could increase the performance have been identified, although we note that the best functions and parameters to use will vary from dataset to dataset, and that the learning mechanism may need to be adjusted depending on the observed 'true' evaluations of congestion.

In our case, median velocity data across the entire network was used, however ideally if we want to be able to measure congestion at the level of

Table 5: Average Spearman correlation over 3 days when predicting congestion from the same day the previous week, where index is calculated based on all 5 regions. Predicting from volume alone would have a correlation of $\rho = -0.781$.

Standard count data

| | WAM | (q) | $I_w$ | (q) | V | (q) | Cons | (q) | Gini | (q) |
|---|---|---|---|---|---|---|---|---|---|---|
| Fitting | -0.825 | 0.422 | -0.748 | -0.008 | -0.746 | 0.209 | -0.137 | 0.426 | -0.714 | 0.443 |
| Prediction | -0.743 | 0.453 | -0.607 | 0.176 | -0.594 | 0.379 | -0.378 | 0.399 | -0.617 | 0.512 |
| Adjusted | -0.770 | -0.764 | -0.852 | **-0.853** | -0.849 | -0.829 | -0.766 | -0.790 | -0.795 | -0.817 |

Counts transformed relative to highest volume over each day

| | WAM | (q) | $I_w$ | (q) | V | (q) | Cons | (q) | Gini | (q) |
|---|---|---|---|---|---|---|---|---|---|---|
| Fitting | -0.819 | 0.422 | -0.048 | 0.256 | -0.001 | 0.278 | 0.107 | 0.379 | 0.076 | 0.342 |
| Prediction | -0.764 | 0.480 | 0.005 | 0.194 | 0.007 | 0.231 | 0.121 | 0.338 | 0.085 | 0.303 |
| Adjusted | -0.781 | -0.778 | -0.721 | -0.753 | -0.720 | -0.755 | -0.765 | -0.788 | -0.753 | -0.794 |

*Bold denotes best prediction result overall.

small regions, we would want data that reflects the traffic status at this local level. Some of our results concerning which metrics performed best could certainly be influenced by this. We also did not experiment with the size or geographical position of regions other than differences in our five example regions.

Since our training data was not exact, we opted for constraints based on consecutive time-steps, however in some cases it may be better to use ordering constraints based on a total ordering of the data, e.g. such that $y_1 > y_2 > \cdots > y_m$ and then we have $\mathcal{P} = \{(1,2),(2,3),(3,4),\ldots,(m-1,m)\}$. We can also use all $m(m-1)/2$ pairwise comparisons. Although in general we would not assume linear correlation or commensurability, it may also be possible to perform numerical fitting based on a transformation of the observed output.

We found that since volume was a fairly robust predictor of congestion, an adjusted index based on inequality and volume could provide better performance than any of the indices alone. Other constructions that allow for more or less influence of volume could also be used, and such parameters could be learned from data using 2-step optimization techniques. The idea of incorporating both volume and spread then gives rise to the potential for density-based averages, which could be explored in a future work. It should also be noted that the traffic volume data itself can be over-estimated, since large trucks etc., can be counted as multiple cars.

Lastly, we highlight that having reliable indices for objectively measuring traffic congestion is crucial for then being able to take steps to reduce it. In real time analysis, measurements across the network can be used to assist traffic monitoring decisions, e.g. changing light sequences, and to inform road-users about the current state of traffic. For retrospective analysis, congestion measurements can be used to evaluate the impact of different strategies or decisions, e.g. in planning for new roads, conducting roadworks, co-ordinating public transport systems and so on.

## 6. Conclusion

We investigated a practical application of inequality and spread measures and proposed methods for learning the weights of such functions from comparison data. We investigated the performance of such techniques when modeling congestion based on counts of traffic passing through multiple intersections throughout a city's road network. We found that although volume and weighted volume were, in general, more reliable than inequality indices, metrics that combined the two provided even better performance.

## Acknowledgement

The authors would like to acknowledge Brisbane City Council and GCS Agile for providing the data used for our experiments.

## References

[1] A review of soft consensus models in a fuzzy environment. *Information Fusion*, 17:4 – 13, 2014.

[2] R. V. Alatalo. Problems in the measurement of evenness in ecology. *Oikos*, 37:199–204, 1981.

[3] J. Alcalde-Unzu and M. Vorsatz. Measuring consensus: concepts, comparisons and properties. In *Consensual Processes, STUDFUZZ*, volume 267, pages 195–211. Springer-Verlag, Berlin, 2011.

[4] O. Aristondo, J.L. García-Lapresta, C. Lasso de la Vega, and R. A. Marques Pereira. The Gini index, the dual decomposition of aggregation functions, and the consistent measurement of inequality. *International Journal of Intelligent Systems*, 27:132–152, 2012.

[5] O. Aristondo, J.L. García-Lapresta, C. Lasso de la Vega, and R. A. Marques Pereira. Classical inequality indices, welfare and illfare functions, and the dual decomposition. *Fuzzy Sets and Systems*, 228:114–136, 2013.

[6] G. Beliakov. Construction of aggregation functions from data using linear programming. *Fuzzy Sets and Systems*, 160:65–75, 2009.

[7] G. Beliakov, H. Bustince, and T. Calvo. *A Practical Guide to Averaging Functions*. Springer, Berlin, New York, 2015.

[8] G. Beliakov, T. Calvo, and S. James. Consensus measures constructed from aggregation functions and fuzzy implications. *Knowledge-Based Systems*, 55:1–8, 2014.

[9] G. Beliakov, M. Gagolewski, and S. James. Penalty-based and other representations of economic inequality. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24:1–23, 2016.

[10] G. Beliakov, S. James, and D. Nimmo. Learning aggregation weights from 3-tuple comparison sets. In *Proceedings of IFSA 2014*, pages 1388–1393. Edmonton, Canada, 2013.

[11] G. Beliakov, S. James, and D. Nimmo. Can indices of ecological evenness be used to measure consensus? In *Proceedings of the 2014 IEEE International Conference on Fuzzy Systems*, pages 1–8. Beijing, China, 2014.

[12] G. Beliakov, S. James, and D. Nimmo. Using aggregation functions to model human judgements of species diversity. *Information Systems*, 306:21–33, 2015.

[13] G. Beliakov, S. James, and D. G. Nimmo. Using aggregation functions to model human judgements of species diversity. *Information Sciences*, 306:21 – 33, 2015.

[14] D. J. Best and D. E. Roberts. Algorithm AS 89: The upper tail probabilities of Spearman's rho. *Applied Statistics*, 24:377–379, 1975.

[15] P. Bloomfield and W.L. Steiger. *Least Absolute Deviations. Theory, Applications and Algorithms*. Birkhauser, Boston, Basel, Stuttgart, 1983.

[16] R. Bosch. *Characterizations of Voting Rules and Consensus Measures, Ph.D. Thesis*. Universiteit van Tilburg, 2006.

[17] F. J. Cabrerizo, S. Alonso, I. J. Pérez, and E. Herrera-Viedma. On consensus measures in fuzzy group decision making. In *Proceedings of MDAI 2008*, pages 86–97, 2008.

[18] S.R. Chakravarty. *Ethical Social Index Numbers*. Springer-Verlag, New York, 1990.

[19] C. Chen. Freeway performance measurement system (pems). *California PATH Research Report*, 2003.

[20] H. Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30:348361, 1920.

[21] E. D'Andrea and F. Marcelloni. Detection of traffic congestion incidents from GPS trace analysis. *Expert Systems with Applications*, 73:43–56, 2017.

[22] B. Dutta. Inequality, poverty and welfare. In K.J. Arrow, A. K. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*, pages 597–633. Elsevier, 2002.

[23] M. Gagolewski. Spread measures and their relation to aggregation functions. *European Journal of Operational Research*, 241(2):469–477, 2015.

[24] J. L. García-Lapresta and D. Pérez-Román. Some measures of consensus generated by distances on weak orders. In *Proceedings of XIV Congreso Español sobre Tecnologías y Lógica fuzzy*, pages 477–483, Cuencas Mineras, 17 –19 September, 2008.

[25] J. L. García-Lapresta and D. Pérez-Román. Measuring consensus in weak orders. In E. Herrera-Viedma, J. L. García-Lapresta, J. Kacprzyk, M. Fedrizzi, H. Nurmi, and S. Zadrozny, editors, *Consensual Processes*, pages 1–21. Springer-Verlag, Berlin, Germany, 2011.

[26] C. Gini. *Variabilità e Mutabilità*. Tipografia di Paolo Cuppini, Bologna, 1912.

[27] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap. *Aggregation Functions*. Cambridge University Press, Cambridge, 2009.

[28] D. Kahle and H. Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013.

[29] A. Mazloumian, N. Geroliminis, and D. Helbing. The spatial variability of vehicle densities as determinant of urban network capacity. *Philosophical Transactions of the Royal Society*, 368:4627–4647, 2010.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[31] Roads and Maritime Agency. Sydney coordinated adaptive traffic system (scats). *Last visited 20 Jan, 2017: http://www.scats.com.au*.

[32] E. H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

[33] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72 – 101, 1904.

[34] C. Taillie. Species equitibility: a comparative approach. In *Ecological diversity in theory and practice*, pages 51–62. Int. Coop. Publ. House, Fairland, Maryland, 1979.

[35] V. Torra and Y. Narukawa. *Modeling Decisions. Information Fusion and Aggregation Operators*. Springer, Berlin, Heidelberg, 2007.

[36] Y. Wang, J. Cao, W. Li, T. Gu, and W. Shi. Exploring traffic congestion correlation from multiple data sources. *Pervasive and Mobile Computing*, http://dx.doi.org/10.1016/j.pmcj.2017.03.015, 2017.

[37] K. Wu, Y. Chen, J. Ma, S. Bai, and X. Tang. Traffic and emissions impact of congestion charging in the central Beijing urban area: A simulation analysis. *Transportation Research Part D*, 51:203–215, 2017.

[38] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. on Systems, Man and Cybernetics*, 18:183–190, 1988.

[39] S. Yang. On feature selection for traffic congestion prediction. *Transportation Research Part C*, 26:160–169, 2013.

[40] S. Yitzhaki. More than a dozen alternative ways of spelling Gini. *Research on Economic Inequality*, 8:13–30, 1998.

[41] X. Zhang, E. Onieva, A. Perallos, E. Osaba, and V. C. S. Lee. Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. *Transportation Research Part C*, 43:127–142, 2014.