An Approach to Sparse, Fine-Grained OD Estimation

Aditya Krishna Menon (Corresponding Author)

National ICT Australia and the Australian National University
7 London Circuit, Canberra, ACT 2612
Mail: Locked Bag 8001, Canberra ACT 260; **Email**: aditya.menon@nicta.com.au
Phone: +612 6267 6293

Chen Cai

National ICT Australia
13 Garden Street, Eveleigh, Sydney, NSW Australia
Mail: Locked Bag 9013, Alexandria NSW 1435; **Email**: chen.cai@nicta.com.au
Phone: +612 9376 2016

Weihong Wang

National ICT Australia
13 Garden Street, Eveleigh, Sydney, NSW Australia
Mail: Locked Bag 9013, Alexandria NSW 1435; **Email**: weihong.wang@nicta.com.au
Phone: +612 8306 0678

Wen Tao

National ICT Australia and the University of New South Wales
13 Garden Street, Eveleigh, Sydney, NSW Australia
Mail: Locked Bag 9013, Alexandria NSW 1435; **Email**: tao.wen@nicta.com.au
Phone: +614 1688 6168

Fang Chen

National ICT Australia and the University of New South Wales
13 Garden Street, Eveleigh, Sydney, NSW Australia
Mail: Locked Bag 9013, Alexandria NSW 1435; **Email**: fang.chen@nicta.com.au
Phone: +612 9376 2101

Word count: 6,200 text + 1 table + 3 figures = 7,200 words

Final submission: November 12th, 2014

Abstract

Given a road network, a fundamental object of interest is the matrix of origin destination (OD) flows. Estimation of this matrix involves at least three sub-problems: (i) determining a suitable set of traffic analysis zones, (ii) the formulation of an optimisation problem to determine the OD matrix, and (iii) a means of evaluating a candidate estimate of the OD matrix. This paper describes a means of addressing each of these concerns using machine learning. We propose to automatically uncover a set of fine-grained traffic analysis zones based on observed link flows. We then employ appropriate regularisation to encourage the estimation of a *sparse* OD matrix. We finally propose to evaluate a candidate OD matrix based on its predictive power on *held out* link flows. Analysis of our approach on a real-world transport network reveals that it uncovers a set of detailed zones, and a corresponding OD matrix that accurately predicts observed link flows.

Keywords: OD estimation, traffic analysis zones, sparsity

1 Problem statement

Consider a directed graph $\mathcal{G}$ representing a road network, where the nodes in the graph represent traffic intersections, and the links represent road segments between intersections. We denote the set of nodes by $\mathcal{N}$ and the set of links by $\mathcal{L}$, and will often write $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. Suppose also that there is a vector $\mathbf{y} \in \mathbb{N}_+^{|\mathcal{L}|}$, representing the count of steady-state traffic flow on each road segment over some time period (e.g. the AM peak). The matrix of origin destination (OD) flows is a fundamental object of interest in the study of the network $\mathcal{G}$ (Ortuzar and Willumnsen, 2011, Chapter 5). In principle, this is a $|\mathcal{N}| \times |\mathcal{N}|$ matrix which represents, for any pair of nodes $(v, v') \in \mathcal{N} \times \mathcal{N}$, the steady-state flow of traffic[1] that begins at v and ends at v'. In practice, one typically focusses on origins and destinations comprising an *aggregation* of nodes[2]; specifically, we identify a special set of *virtual nodes* $\mathcal{Z}$, which represent the focal points of certain *traffic analysis zones*, and consider the $|\mathcal{Z}| \times |\mathcal{Z}|$ OD matrix, which we write as $\mathbf{X}$. Each virtual node is connected via a number of *virtual links* to nodes in $\mathcal{N}$. The presence of such a link originating from z represents that the corresponding non-virtual node belongs to zone z; each $z \in \mathcal{Z}$ thus represents an aggregation of nodes in the original graph.

The OD matrix is a valuable tool for understanding and forecasting usage patterns of a network. Given the OD matrix, one can then make forecasts about traffic flows on a *different network* $\mathcal{G}' = (\mathcal{N}', \mathcal{L}')$, under the assumption that the networks $\mathcal{G}$ and $\mathcal{G}'$ possess commensurate OD flows. For example, the network $\mathcal{G}'$ might be identical to $\mathcal{G}$, except that certain links are removed; the forecast flows may then be used to assess the impact this change has on the network. The predicted flows may be generated by any *route assignment* model, such as for example one based on a user equilibrium assumption (Sheffi, 1985, Chapter 3).

This paper is concerned with the *OD estimation problem*. Here, the aim is to recover $\mathbf{X}$ given the topology of the road network $\mathcal{G}$, observed link flows $\mathbf{y}$, and the definition of the traffic analysis zones $\mathcal{Z}$. Any attempt at OD estimation faces several entwined questions:

- how does one define the traffic analysis zones $\mathcal{Z}$? As the choice of $\mathcal{Z}$ defines the precise pairwise flows we are interested in estimating, it plays a crucial role in determining whether the resulting OD matrix can be reliably estimated, and whether it is useful for analysis and forecasting.
- given $\mathcal{Z}$ and link flows $\mathbf{y}$, how does one estimate the OD matrix $\mathbf{X}$? The OD matrix can be understood as the solution to a potentially ill-posed linear system. Its estimation thus requires some means of choosing amongst potentially multiple candidate OD matrices.
- given an estimate of the OD matrix, $\hat{\mathbf{X}}$, how does one evaluate its efficacy? As there is typically no direct ground truth for the OD matrix, any analysis of the quality of its

[1] More generally, one may be interested in time-varying OD matrices. While the topic of considerable research in its own right (Cascetta et al., 2013), we do not consider this problem here.

[2] While virtual nodes aggregate the original nodes in the graph, one still retains the original nodes for all subsequent modelling and analysis. This is because the original nodes may be used as intermediate nodes for travelling from one zone to another.

estimate must rely on auxiliary measures.

In this paper, we explore techniques to answer all three questions. In a nutshell, we propose:

- the *automated design* of *fine-grained* traffic analysis zones, based on the intuition that a good set of analysis zones has minimal intra-zonal flow (i.e. flows that begins and ends at nodes within the same traffic analysis zone);
- an OD estimation procedure that encourages the estimation of *sparse* OD matrices, which, in addition to mitigating issues of ill-posedness, are generally interpretable;
- the use of *held-out flow predictions* to evaluate efficacy of an estimated OD matrix $\mathbf{X}$, by viewing OD estimation as a type of general regression problem.

We evaluate our approach on a real-world network, and find that we can discover an intuitive zoning of the network, and learn an OD matrix that reliably predicts link flows.

This paper is organised as follows. In §2, we discuss the above three challenges in more detail, and describe prior work in the literature on addressing these challenges. Then, in §3, we detail the elements of our solution, which attempt to employ machine learning to aid in solving the estimation problem. We then evaluate our method on a real-world network in §4. We conclude in §5 with some discussion on areas for future research.

2 Challenges of OD estimation

In this section, we review the challenges involved in each of the three items we described earlier. We also discuss existing work that we are aware of to deal with these challenges.

2.1 Zoning

To appreciate the challenges inherent in the design of traffic analysis zones $\mathcal{Z}$, it is first worth noting the implications of two extreme choices of $\mathcal{Z}$. Ideally one would like to set $\mathcal{Z} = \mathcal{N}$, so that the OD matrix comprises flows between each pair of intersections. The drawback of this choice is that, as shall be made precise in the next section, this potentially leads to a highly ill-posed estimation for the OD matrix, as one needs to solve for $|\mathcal{N}|^2$ unknowns given only $|\mathcal{L}|$ equations. A computational drawback is that at increasing level of granularity, estimating $|\mathcal{N}|^2$ parameters may be infeasible on networks where $|\mathcal{N}| \approx 1000$.

Conversely, by choosing $|\mathcal{Z}|$ to be small, one mitigates ill-posedness, and there are no computational barriers. However, this comes at a significant expense: *intra-zonal flows* – i.e. flows between any pair of nodes v, v' that are in the same zone – are ignored. Assuming that the OD matrix is to be used for forecasting under changes to the network, it is likely unacceptable to ignore the impact of any change to high-volume links.

Between these two extremes, then, one faces a tradeoff between statistical and predictive precision. The precise choice as to this tradeoff is often left as a specification for a domain expert (Ortuzar and Willumnsen, 2011). There have been alternate proposals that attempt to make the procedure more automated. A notable example is the work of Martínez et al. (2009), who propose an optimisation framework that takes into account several desiderata for the design of analysis zones, including the minimisation of intra-zonal flows as mentioned, but also the geographic contiguity of the resulting partition of the road network. The framework relies on the availability of a sufficiently detailed prior OD matrix derived from survey data, however.

2.2 OD estimation

Suppose we have fixed our traffic analysis zones $\mathcal{Z}$, so that the OD matrix $\mathbf{X} \in \mathbb{N}_+^{|\mathcal{Z}|\times|\mathcal{Z}|}$. For convenience, we shall interchangeably refer to the OD matrix by $\mathbf{X}$ and its vectorised form, $\mathbf{x} = \mathrm{vec}(\mathbf{X}) \in \mathbb{N}_+^{|\mathcal{Z}|^2\times 1}$. There are roughly three approaches to estimating the OD matrix (Cascetta, 1984):

- The simplest is *direct sample estimation*, wherein surveys or interviews are conducted to determine common origin-destination pairs for individuals. Aggregating these responses gives estimates of the OD matrix cells.
- Survey information is often incomplete, and thus may provide no (or highly biased) information about certain OD pairs. Inferences can nonetheless be made by performing *model estimation*, wherein a particular model is assumed to relate OD flow to several explanatory variables, such as the mean income of residents with a zone. A well-known instance of this approach is the *gravity model* (Ortuzar and Willumnsen, 2011, pg. 182), (Zhang et al., 2003).
- Survey information is often based on small samples, and thus unreliable. A richer class of techniques are those based on *estimation from loop counts* (Willumsen, 1981). Loop counts embed information about OD and route choices, and are typically more plentiful and reliable than survey data. These are sometimes referred to as structured and unstructured methods, or parameter calibration and matrix estimation methods respectively (Tamin and Willumsen, 1989).

We will focus on the latter class of methods, noting that they may be extended to exploit survey information if it is available. Examples of methods that exploit link counts include those based on maximum entropy modelling (Van-Zuylen and Willumsen, 1980), maximum likelihood estimation (Spiess, 1987), and Bayesian inference (Maher, 1983). A basic fact that underlies these approaches is that the OD matrix $\mathbf{x}$ is related to the link flows $\mathbf{y}$ via the *flow-conservation* equation,

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \tag{1}$$

where $\mathbf{A} \in [0,1]^{|\mathcal{L}|\times|\mathcal{Z}|^2}$ is the *assignment map*, whose entries denote the probability of a particular link being used for travel between an OD pair. Estimating the OD matrix is thus

equivalent to solving this linear system. There are at least two challenges with doing so. First, as noted in the previous section, the linear system is ostensibly strongly ill-posed or undetermined: we have $|\mathcal{Z}|^2$ unknowns and $|\mathcal{L}|$ equations. This means there may not be a unique $\mathbf{x}$ satisfying Equation 1. Second, in congested scenarios, the assignment map $\mathbf{A}$ itself depends on the optimal OD matrix, for example based on an equilibrium assignment. Therefore, the design matrix must *itself* be estimated from the link flows.

A strategy to mitigate ill-posedness is to inject some prior or domain knowledge into the estimation problem. Typically, this is done by relying on prior OD estimates collected e.g. from a survey. Suppose $\mathbf{x}^{\text{old}}$ denotes this prior estimate of the OD matrix. Then, the generalised least squares estimator (Cascetta and Nguyen, 1988) aims to find

$$\min_{\mathbf{x} \succeq \mathbf{0}} (\mathbf{A}\mathbf{x} - \mathbf{y})^T \mathbf{W}^{-1} (\mathbf{A}\mathbf{x} - \mathbf{y}) + (\mathbf{x} - \mathbf{x}^{\text{old}})^T \mathbf{V}^{-1} (\mathbf{x} - \mathbf{x}^{\text{old}}), \tag{2}$$

where $\mathbf{W}, \mathbf{V}$ are appropriate covariance matrices for the distributions of the prior $\Pr[\mathbf{x}]$ and likelihood $\Pr[\mathbf{y}|\mathbf{x}; \mathbf{A}]$ respectively. If one ignores the nonnegativity constraint, the above admits a closed-form solution. When interpreted from a Bayesian perspective, one may also obtain an estimate of the posterior covariance of the OD matrix (Cascetta and Nguyen, 1988).

A strategy to overcome the fact that $\mathbf{A}$ itself depends on $\mathbf{X}$ is to cast the problem in the framework of bilevel programming (Yang et al., 1992). Practically, what this amounts to is the alternating solution of the linear system with respect to $\mathbf{X}$, and an appropriate objective for $\mathbf{A}$ that loads $\mathbf{X}$ onto the network. The latter may be cast as an convex problem for a range of equilibrium based assignments, such as a deterministic user equilibrium (Sheffi, 1985).

Some comments on the ill-posedness of Equation 1 are appropriate. First, ill-posedness is guaranteed if there is some $\mathbf{x} \succeq \mathbf{0}$ such that $\mathbf{A}\mathbf{x} = \mathbf{0}$; this is because $\mathbf{x}$ can then be added onto any candidate solution to Equation 1 without affecting the flow estimates. Second, the non negativity constraint on $\mathbf{x}$ cannot be ignored in assessing the uniqueness of the linear system: Wang and Tang (2009); Wang et al. (2011) have shown that when there exists a sparse solution to the system, it may be the *only* solution. Bierlaire (2002) proposed a surrogate measure of the degree of ill-posedness of the system, taking this fact into account. Third, a pleasant consequence of using a prior OD matrix $\mathbf{x}^{\text{old}}$ is that the objective corresponding to e.g. Equation 2 is strictly convex (assuming the diagonal entries of $\mathbf{V}$ are positive), meaning that there will in fact be a unique solution: amongst all OD matrices that have the same predicted link flows, we seek the one that is closest to the given prior OD matrix. Fourth, the system is potentially well-posed if one considers correlations in flows across *multiple* days. For example, Vardi (1996) showed that under mild assumptions, the OD matrix is identifiable for Poission models. Hazelton (2001) showed that second-order information present due to temporal trends may also induce identifiability.

2.3 Evaluation

Perhaps the ideal means of assessing the quality of OD estimates is in the ability to accurately forecast traffic flows under a changed network $\mathcal{G}'$ with commensurate demands; however, ma-

jor network changes of this type are not common, making it difficult to acquire the necessary data. Consequently, much prior work on OD estimation evaluates the efficacy of the estimation procedure by applying it to a synthetic network where the ground-truth OD is known. While sensible, it is of interest to be able to compare different OD estimates on a real network where the ground-truth is of course unknown. The challenge is determining a suitable auxiliary measure of quality. We are not aware of much work that directly addresses this issue. A natural idea is to assess its predictive power in forecasting link flows at future time periods, but the ill-posedness issue arises: suppose the flow-conservation equation (Equation 1) is ill-posed, so that $\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2$ for some $\mathbf{x}_1 \neq \mathbf{x}_2$. Then several OD matrices will yield exactly the same link flows. An alternative is to resort to interpretability, but this may be difficult to achieve with more fine-grained estimates.

3 Methodology

We are now in a position to describe our methodology, beginning with our scheme for the automated design of traffic analysis zones.

3.1 Zoning

Our design of traffic analysis zones hinges on a few simple observations that are worth explicating. First, as noted previously, while a per-node zoning scheme ($\mathcal{Z} = \mathcal{N}$) poses statistical and computational challenges, it is the gold standard in terms of modelling capability; any other zoning scheme necessarily contains no more information, by virtue of aggregation. Therefore, the per-node zoning serves as a useful starting point from which to begin any attempt at zoning.

Second, it is not necessary to associate every node $v \in \mathcal{N}$ with a traffic analysis zone. Crucially, this does not preclude any links involving them being a component of the paths between some origin-destination pair. Consider a network with a single source and sink node, and several intermediate nodes. Here, only the source and sink need be designated as belonging to a zone; the other nodes may safely be omitted from the zoning analysis.

Third, as our primary interest is in the use of the OD matrix to forecast the effect of changes to the network, our primary concern is the minimisation of intra-zonal flow[3]. This means that, minimally, we require that for the *high volume* links at some suitable threshold τ,

$$\mathcal{L}_\tau^{\mathrm{HV}} = \{e \in \mathcal{L} : \mathbf{y}_e \geq \tau\},$$

the majority of node pairs relying on these links must not be assigned to the same zone.

Based on this, we consider a simple zoning scheme that is derived from an initial per-node zoning. Performing a basic route assignment with such a zoning allows us to study the likely

[3]There may of course be other desiderata or constraints when constructing a zoning, for example, respecting suburban or demographic divisions. We do not directly consider these in this paper, as our goal is simply to be able to accurately forecast flows.

explanations for traffic on each high volume link in $\mathcal{L}_\tau^{\mathrm{HV}}$. Our intuition is that if a pair of nodes is seen to rely on *several* such high volume links, then they likely[4] represent areas of interest. Thus, assigning each of these nodes into a separate zone encourages appropriate assignment of flow for the high volume links. Roughly, the resulting zones represent dominant regions of in- and/or out-flow.

In detail, we perform zoning as follows.

1. First, employ a per-node zoning scheme, and assign traffic onto the corresponding network using an appropriate route assignment algorithm. Following our earlier discussion, one simple option is to employ a deterministic user-equilibrium assignment. The result of the route assignment is an assignment map $\mathbf{A} \in [0,1]^{|\mathcal{L}|\times|\mathcal{N}|^2}$.

2. The matrix $\mathbf{A}$ encodes, for each link, the OD pairs that rely on it for travel. Given a link $e \in \mathcal{L}$, and a suitable parameter $\delta \in [0,1]$, let

$$\mathcal{P}_\delta(e) = \{(v,v') \in \mathcal{N}\times\mathcal{N} : \mathbf{A}_{e(v,v')} \geq \delta\}$$

   denote the pairs of nodes that rely on that link e for travel with probability at least δ.

3. As we have the observed link flows $\mathbf{y}$, we can compute $\mathcal{L}_\tau^{\mathrm{HV}}$ for an appropriate threshold τ. We then count, for each pair of nodes (v,v'), the number of high volume links that are relied on for travel with probability bigger than δ:

$$c(v,v') = \sum_{e\in\mathcal{L}_\tau^{\mathrm{HV}}} [\![(v,v') \in \mathcal{P}_\delta(e)]\!]. \tag{3}$$

4. We finally ensure that for each high volume link $e \in \mathcal{L}_\tau^{\mathrm{HV}}$, there is at least one OD pair in $\mathcal{P}_e$ that is selected for our final zoning. We do this by scanning through all the high volume links, and collecting the OD pair in $\mathcal{P}_e$ which has the highest count given by Equation 3. That is, beginning with an empty set of OD pairs $\mathcal{P} = \emptyset$, we update for each e via

$$\mathcal{P} \leftarrow \mathcal{P} \cup \left\{ \operatorname*{Argmax}_{(v,v')\in\mathcal{P}_\delta(e)} c(v,v') \right\}.$$

   We then simply assign each node in $\mathcal{P}$ to a separate zone.

Some comments are in order. First, the route assignment in step (1) requires an initial OD estimate; if this is present, it may be used. In our experiments, we did not have access to a per-node OD matrix. Therefore, we simply employed a uniform OD matrix over the sites; while not ideal, we found this to provide good results on the real-world network we have experimented with. Second, the choice of τ determines our tolerance for potentially missing

[4]We say "likely" because our routing assignment with a per-node zoning will have to rely on a rough estimate of the OD matrix at this granularity. Bias in this matrix may be reflected in the corresponding routing.

out flow on certain links. The choice of δ similarly determines how crucial a particular OD pair is to explaining traffic on a link. While there is the potential for automated setting of these parameters in turn, we simply experimented with a few values to determine which minimised the amount of intra-zonal flow, as well as the number of zones itself.

3.2 OD estimation

Having defined a set of fine-grained traffic analysis zones, we turn to the problem of estimating the OD matrix $\mathbf{x}$. Our basic idea is that for a fine-grained zoning, the underlying OD matrix should be *sparse*: for most OD pairs, there should be exactly zero flow between that pair. This is because our intuition is that the traffic observed in the network should largely be the result of the travel between a small subset of the $|\mathcal{Z}|^2$ candidate pairs, which e.g. reflects that during the AM peak, we expect there to be a few popular destinations (corresponding to office locations, parking lots, and so on), with most other zones seeing minimal in-flow. We may similarly expect most origins of flow in a business district to come from the boundaries of the network (corresponding to suburbs and residential areas).

The question then is how we can achieve the sparsity in the OD matrix. Following the GLS objective (Equation 2), we will consider an objective of the form

$$\min_{\mathbf{x} \succeq \mathbf{0}} (\mathbf{A}\mathbf{x} - \mathbf{y})^T \mathbf{W}^{-1} (\mathbf{A}\mathbf{x} - \mathbf{y}) + \Omega(\mathbf{X}). \tag{4}$$

Here, $\mathbf{W}$ is a diagonal matrix whose entries are of the form

$$(\forall e \in \mathcal{L})\, \mathbf{W}_{ee} = \mathbf{y}_e^{\beta}$$

for some $\beta \in \mathbb{R}_+$. This represents non-isotropic noise in the observed link flows, with higher link flows (corresponding to more heavily used roads) subject to higher errors. When $\beta = 1$, this is seen to mimic a Poisson model. Further, Ω denotes our generic regulariser, which is of the form

$$\Omega(\mathbf{X}) = \lambda_1 ||\mathbf{x}||_1 + \lambda_2 (\mathbf{x} - \mathbf{x}^{\text{old}})^T \mathbf{V}^{-1} (\mathbf{x} - \mathbf{x}^{\text{old}}). \tag{5}$$

Here, $\mathbf{x}^{\text{old}}$ is a prior OD matrix, derived for example from survey data. The matrix $\mathbf{V}$ is diagonal matrix, with entries of the form

$$(\forall z \in \mathcal{Z})\, \mathbf{V}_{zz'} = (\mathbf{X}^{\text{old}}_{zz'})^{\alpha}$$

for some $\alpha \in \mathbb{R}_+$. We shall refer to the term $(\mathbf{x} - \mathbf{x}^{\text{old}})^T \mathbf{V}^{-1} (\mathbf{x} - \mathbf{x}^{\text{old}})$ as an ℓ_2 *regulariser*, as it is a weighted version of the ℓ_2 norm of $\mathbf{x} - \mathbf{x}^{\text{old}}$.

The term $||\mathbf{x}||_1 = \sum_{i,j=1}^{|\mathcal{Z}|} |\mathbf{X}_{ij}|$ encodes the belief that the true OD matrix is sparse. The intuition for such a term inducing sparsity follows by interpreting it as a convex relaxation to the ℓ_0 "norm", which is exactly the number of non zeros in $\mathbf{x}$. We shall refer to this term as an ℓ_1 *regulariser*. The use of ℓ_1 regularisation to discover sparse solutions has seen wide use in compressed sensing (Donoho, 2006) and in applications of the Lasso algorithm (Tibshirani, 1996).

It is worth noting that for the case where $\lambda_1 = 0$, our objective exactly matches the classical one considered by generalised least squares (GLS) estimators of the OD, or equivalently appropriately set-up probabilistic models for the likelihood and prior. For $\lambda_1 > 0, \lambda_2 > 0$, the regulariser can be seen as a variant of the elastic net (Zou and Hastie, 2005).

The idea of using ℓ_1 regularisation for problems involving OD estimation is not new. It has been proposed previously in at least Chawla et al. (2012); Mardani and Giannakis (2013); Sanandaji and Varaiya (2014), albeit with slightly different contexts and motivations: the former two works are concerned with robustness to anomalies, and the latter with sparsity in path flows. However, these works do not explicitly consider constraint that our OD matrix elements must be nonnegative. While this constraint seems innocuous, it has some important implications. First, it has been recently observed that a non negativity constraint *by itself* may induce sparse solutions (Slawski and Hein, 2012; Meinshausen, 2013). The reason for this is a certain "self-regularising" property of the design matrix, which in our case is the assignment map $\mathbf{A}$. Indeed, Meinshausen (2013) observed in synthetic experiments on network tomography that non negativity is robust as a regulariser. This fact suggests that it is not *a priori* obvious that the flow conservation equation (Equation 1) is ill-posed; the non negativity constraint may sufficiently regularise the system so as to guarantee uniqueness (Wang and Tang, 2009). Minimally, by acting as a regulariser, it may encourage the estimation of OD matrices that are useful predictors under changes to the network.

Second, the non-negativity of $\mathbf{x}$ allows us to simplify the regulariser to

$$\Omega(\mathbf{X}) = \lambda_1 \mathbf{1}^T \mathbf{x} + \lambda_2 ||\mathbf{x} - \mathbf{x}^{\text{old}}||_2^2.$$

The objective now becomes differentiable everywhere, allowing for the easy application of gradient-based methods. We experimented with the LBFGS-B optimiser (Zhu et al., 1997), which performs quasi-Newton minimisation while respecting the constraint $\{\mathbf{x} \succeq 0\}$. Other approaches are of course possible, including the use of the general purpose quadratic programming solvers, or more general convex optimisation solvers such as the CVX toolbox (Grant and Boyd, 2014, 2008).

3.3 Evaluation

Treating OD estimation for fixed $\mathbf{A}$ as a type of regression problem, a natural strategy to evaluate an OD matrix is based on prediction on flows that are *held-out* during the estimation procedure. This follows the general procedure for evaluating any model estimated from data (Hastie et al., 2009, Chapter 7). The idea is as follows:

1. partition the set of links $\mathcal{L}$ into two sets, $\mathcal{L}_1$ and $\mathcal{L}_2$
2. estimate the OD matrix based *only on link flows from* $\mathcal{L}_1$
3. evaluate the predictive performance *only on link flows from* $\mathcal{L}_2$.

This process may be repeated many times. The average of the performance computed in step (3) may be taken as an estimate of the predictive power of the estimated OD matrix $\hat{\mathbf{x}}$.

Formally, step (2) is equivalent to solving our objective in Equation 4 with the vector $\tilde{\mathbf{y}}$, defined as

$$\tilde{\mathbf{y}}_e = \begin{cases} \mathbf{y}_e & \text{if } e \in \mathcal{L}_1 \\ 0 & \text{else,} \end{cases}$$

and the matrix $\tilde{\mathbf{A}}$, defined as

$$\tilde{\mathbf{A}}_{ep} = \begin{cases} \mathbf{A}_{ep} & \text{if } e \in \mathcal{L}_1 \\ 0 & \text{else.} \end{cases}$$

Put plainly, we simply ignore the links in $\mathcal{L}_2$ when estimating the OD matrix $\mathbf{x}$. Once we have an estimate $\hat{\mathbf{x}}$, we may of course compute the predicted flows on links in $\mathcal{L}_2$ via the full assignment map $\mathbf{A}$. Step (3) then requires that we summarise the fidelity of the predictions on these links alone.

A few comments are in order. First, the partitioning of the links into the two sets is independent of the zoning procedure. A simple strategy is to just perform a random partition. Second, it is essential that step (3) operate only on links in $\mathcal{L}_2$: as the links in $\mathcal{L}_1$ were used to estimate $\hat{\mathbf{x}}$, the performance on these links is a highly biased estimate of the true predictive performance of $\hat{\mathbf{x}}$. Third, the splitting is only performed for the OD estimation. It does *not* mean that the links are physically removed, and in particular, they are still a part of the assignment procedure employed to find the assignment map $\mathbf{A}$. The value of this split is that we can then evaluate the predictive performance of $\mathbf{X}$ on the second, or *held-out*, set of links.

A subtlety is that because of the risk of ill-posedness, it may well be that there are multiple OD matrices that have identical performance on held-out links as well. However, as we encourage agreement with the prior OD matrix, $\mathbf{x}^{\text{old}}$, and further regularise the matrix to possess sparsity (which is an application of our domain knowledge), we believe that good held-out performance of such a regularised estimate is indicative of reliable estimation. In our experiments, we shall further study the interpretability of the matrix, as well as perform a case-study on real link flows.

4 Evaluation on real-world traffic network

We now present results confirming the efficacy of our approach.

4.1 Description of data

We conduct experiments on traffic counts obtained for an urban area during a two-week period in 2012. The data comprises observations for 155 intersections and 310 road segments connecting them, during the period of 7AM - 10AM.

One challenge with the data is the lack of a suitable prior OD matrix. Existing public survey data only provides a limited number of samples, which is insufficiently reliable to use as a benchmark. In our experience, even simple smoothing models, such as the gravity model, do not offer significant improvement in the performance of these estimates. Therefore, we chose to use a naïve *uniform* prior $\mathbf{x}^{\text{old}}$ as our prior OD matrix. All entries in this matrix are equal to $\frac{T}{|\mathcal{Z}|^2}$, with T being the total number of trips in the network. The number T was chosen based on domain knowledge; this initial guess may of course be refined with techniques such as bilevel programming.

4.2 Experimental aims

In brief, the aims of our experimental study are:

- To assess the quality of the zoning induced by our approach described in §3.1. In particular, we assess how successful this approach is in attaining our goal of minimising intra-zonal flow, and how it compares to a manually defined zoning based on domain knowledge.
- To determine the impact that enforcing of the nonnegativity constraint has on the quality of the OD estimates, in terms of predictive power, as well as sparsity of the resulting solution. We further aim to assess how adding ℓ_2 and ℓ_1 regularisation affect these properties.
- To determine the value of assessing OD estimates using held-out flows, by for example demonstrating that it rejects estimates that grossly overfit to the observed link flows.

We thus vary three knobs, and report results for each ensuing combination:

- For the *Zoning*, we report results on both a manual, coarse-grained zoning determined by a domain expert, and the fine-grained zoning determined by our approach (§3.1).
- For the *Learner*, we either simply report the prior OD matrix ("None"), minimise the objective in Equation 4 without a nonnegativity constraint ("GLS"), or minimise the objective in Equation 4 with the non negativity constraint ("NN-GLS"). For GLS, the final solution is thresholded at 0.
- For the *Regulariser* Ω, we test no regularisation ($\lambda_1 = \lambda_2 = 0$ in Equation 5), ℓ_2 regularisation to the prior OD matrix ($\lambda_1 = 0, \lambda_2 > 0$), ℓ_1 regularisation to induce sparsity ($\lambda_1 > 0, \lambda_2 = 0$), and a combination of ℓ_1 and ℓ_2 regularisation ($\lambda_1, \lambda_2 > 0$).

We now describe the results of our zoning scheme.

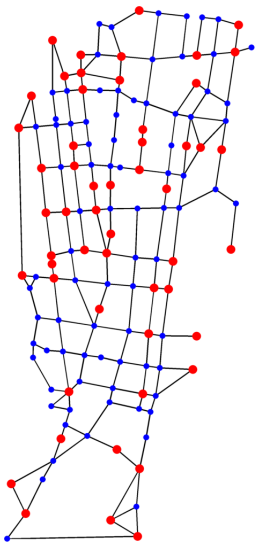


FIGURE 1: Visualisation of results automated zoning procedure. The red (large) dots denote nodes assigned to a separate zone, and the blue (small) dots denote all other nodes. (The node layout corresponds to geographic location. Precise coordinates are omitted for data confidentiality reasons.)

4.3 Zoning results

The manual zoning of the network performed by a domain expert comprises a total of 22 zones. (We call this zoning "coarse-grained" to contrast it with the automated zoning we shall subsequently investigate.) The design of these zones is guided by geographic contiguity, and domain knowledge as to certain popular regions of interest in the network. While the result is an intuitive partition of the network, it suffers from the problem of missing the flows on certain high volume links. This is illustrated in Figure 2a, which compares the predicted and true flows on each link.

We applied our zoning procedure specified in §3.1, with parameters $\tau = 1000$ and $\delta = 1$. This gave a reasonable tradeoff between maximising inter-zonal flow while minimising the number of zones returned. Our procedure resulted in a total of 42 zones, which is only moderately higher than the coarse zoning (though the quadratic growth in OD pairs means it corresponds to 400 extra parameters to be estimated). Recall that our zones comprise individual nodes in the network. Figure 1 shows an overlay of these "node zones" (red) and the standard nodes (blue). This evinces one intuitive property of the zoning, namely, they provide a good coverage of the boundary of the network. We shall now see that the zoning also helps improve prediction of link flows by OD estimation.

4.4 Held-out prediction results

We now report results on the quality of various OD estimation schemes. Following our discussion in §3.3, we report held-out link flow performance of the OD estimates derived from

each method. As performance measures, we report the RMSE, MAE, and Spearman's ρ of our predicted link flows $\hat{\mathbf{y}}$ against the true link flows $\mathbf{y}$. These are defined as

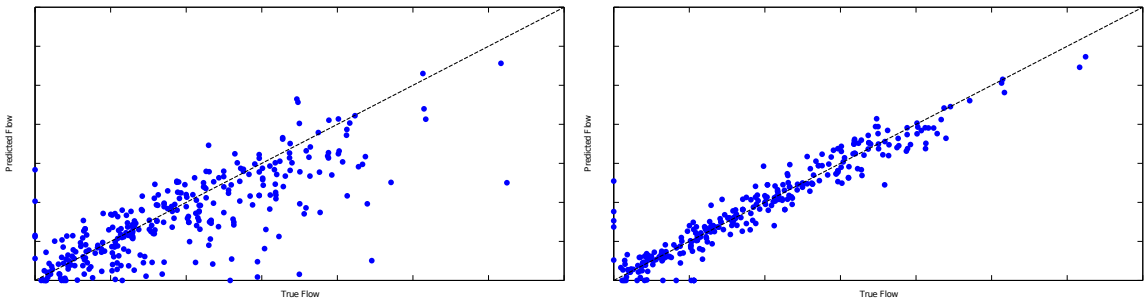
$$\begin{aligned}
\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) &= \sqrt{\frac{1}{|\mathcal{L}|} \sum_e (\mathbf{y}_e - \hat{\mathbf{y}}_e)^2} \\
\mathrm{MAE}(\mathbf{y}, \hat{\mathbf{y}}) &= \frac{1}{|\mathcal{L}|} \sum_e |\mathbf{y}_e - \hat{\mathbf{y}}_e| \\
\rho(\mathbf{y}, \hat{\mathbf{y}}) &= 1 - 6 \frac{\sum_e (\mathbf{r}_e - \hat{\mathbf{r}}_e)^2}{|\mathcal{L}|(|\mathcal{L}|^2 - 1)},
\end{aligned}$$

where $\mathbf{r}, \hat{\mathbf{r}}$ represent the ranks of the links according to the flows $\mathbf{y}, \hat{\mathbf{y}}$. We normalise the RMSE and MAE metrics based on that of the trivial baseline $\bar{\mathbf{y}}$ which predicts the mean and median respectively of the observed link flows, i.e. we report $\frac{\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}})}{\mathrm{RMSE}(\mathbf{y}, \bar{\mathbf{y}})}$. (This is akin to the standard R^2 coefficient of determination, which reports normalised mean squared errors.) Any method must thus attain a normalised score of less than 1 to be considered useful. We call these normalised scores the NRMSE and NMAE respectively.

Table 1 summarises these performance measures from 5 independent trials of partitioning the links. Overall, we find that

- simply relying on the prior OD matrix performs poorly in terms of RMSE and MAE, indicating the value of doing some form of optimisation based on link flows.
- our fine-grained zoning results in superior held-out predictions compared to the manually defined, coarse-grained zoning. This is unsurprising, as the latter is simply unable to offer reasonable predictions for intra-zonal flows. (Observe that when the Learner is "None", it is expected for the fine-grained zoning to perform worse than the coarse-grained counterpart, as a uniform distribution in the former is likely highly biased.)
- imposing a nonnegative constraint on the OD matrix during estimation has a non-trivial impact on performance: we find that NN-GLS outperforms plain GLS for both the coarse- and fine-grained zoning, when both do not employ any additional regulariser.
- NN-GLS by itself is competitive with GLS and ℓ_1 regularisation for the fine-grained zoning. We shall subsequently see that this is also true in terms of the sparsity of the solutions.
- ℓ_2 and ℓ_1 regularisation generally improve performance when used with GLS, with ℓ_2 being slightly more useful, by virtue of shrinking towards prior estimates of the OD. Their combination yields commensurate performance to either regulariser individually. Employing ℓ_1 regularisation has the advantage of improving sparsity of the resulting solution, as we shall see.

Figure 2 shows a scatterplot of the predicted versus true link flows with both the coarse- and fine-grained zoning after performing NN-GLS with ℓ_1, ℓ_2 regularisation. As expected, the fine-grained zoning results in a better fit to the link flows. Also of interest is the fact that there are



(a) Coarse-grained zoning. **(b)** Fine-grained zoning.

FIGURE 2: Scatter plot of predicted versus true link flows. (The axes' scales are unlabelled for data confidentiality reasons.)

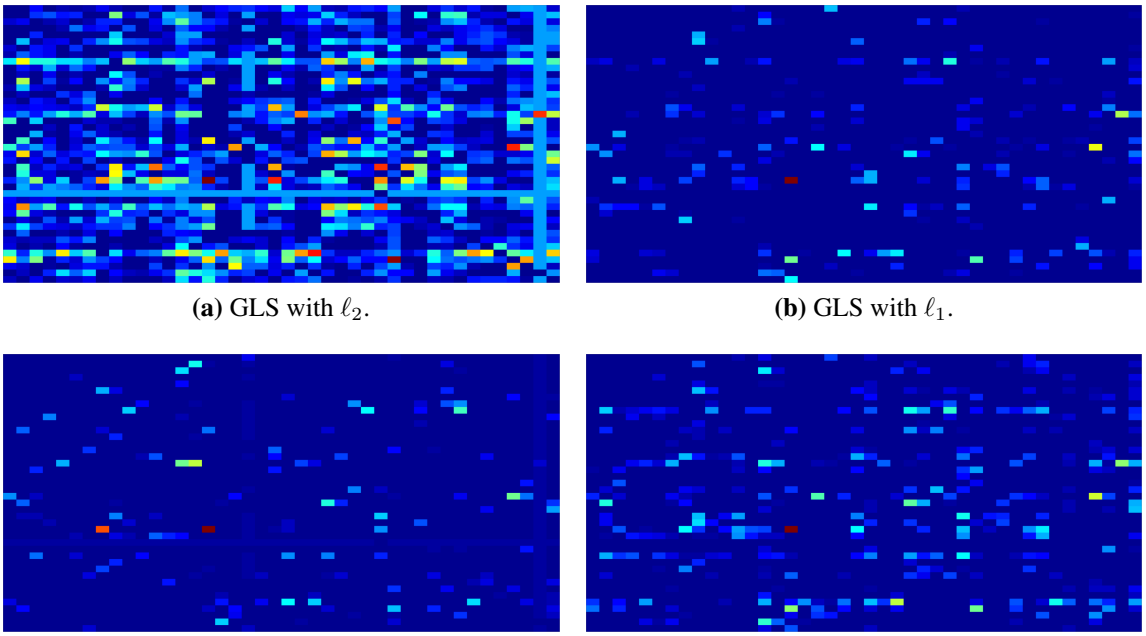
certain links with a moderate amount of flow (≈ 2500) that are missed by the coarse-grained zoning; these are a result of intra-zonal flows that cannot be captured by the manually defined zones.

It is of interest to study the sparsity of the estimates resulting from the various methods. For the fine-grained zoning, we find that GLS with ℓ_2 regularisation, while accurate in terms of predictive power, has 73% of entries with flow greater than 0. By contrast, GLS with ℓ_1 regularisation reduces this to 17%. Of interest then is that NN-GLS, with no regularisation, produces an even sparser OD matrix, with only 14% of entries greater than 1; this is consistent with the theoretical findings of (Slawski and Hein, 2012; Meinshausen, 2013) that non negativity may by itself induce sparsity. Adding ℓ_1 and ℓ_2 brings the sparsity level up to 22%, which is expected as the ℓ_2 regularisation encourages estimates towards the (nonzero) prior OD. Figure 3 visualises the OD matrix for each of these approaches, where each figure is simply a heatmap of the estimated flows between each OD pair. We see that in general, some dominant entries are discovered by all approaches. However, GLS with ℓ_2 regularisation is notable in inducing many small, but non-zero estimates.

5 Conclusion

This work proposed a strategy for the estimation of a sparse, fine-grained OD matrix. In particular, we proposed a scheme for constructing a fine-grained zoning, based on the intuition that we wish to consider OD pairs that explain traffic on high-volume links. We then discussed how to encourage the estimation of a sparse OD matrix, using ℓ_1 regularisation. We also noted the non-trivial role that nonnegativity may play in the sparsity of OD estimates. We finally discussed how held-out link flow prediction can be used to assess the quality of OD estimates. Experimental results on a real-world network show encouraging results for our approach.

There are several avenues for future work. One is to do with the use of bilevel programming to estimate the OD matrix, which we expect will improve performance. A distinct approach



(a) GLS with ℓ_2.

(b) GLS with ℓ_1.

(c) NN-GLS.

(d) NN-GLS with ℓ_1, ℓ_2.

FIGURE 3: OD matrix estimates for various approaches. (Best viewed in colour.)

is to directly estimate path flows. This brings a different set of advantages and disadvantages: one the one hand, one does not need to rely on an alternating optimisation, but on the other hand, one has to estimate a potentially exponential number of variables. The latter fact makes sparsity inducing regularisers a natural candidate, and indeed this has been explored in very recent work (Sanandaji and Varaiya, 2014). It is of interest to see whether the trace norm regulariser may also be used to improve results, as explored in (Mardani and Giannakis, 2013).

Acknowledgements

We thank the Roads and Maritime Services (RMS) for providing data. This work was supported by NICTA. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

Bierlaire, M. (2002). The total demand scale: a new measure of quality for static and dynamic origin-destination trip tables. *Transportation Research Part B: Methodological*, 36(9):837 – 850.

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transportation Research Part B: Methodological*, 18(4-5):289–299.

Cascetta, E. and Nguyen, S. (1988). A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 22(6):437–455.

Cascetta, E., Papola, A., Marzano, V., Simonelli, F., and Vitiello, I. (2013). Quasi-dynamic estimation of OD flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transportation Research Part B: Methodological*, 55(0):171 – 187.

Chawla, S., Zheng, Y., and Hu, J. (2012). Inferring the root cause in road traffic anomalies. In Zaki, M. J., Siebes, A., Yu, J. X., Goethals, B., Webb, G. I., and Wu, X., editors, *ICDM*, pages 141–150. IEEE Computer Society.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited. http://stanford.edu/~boyd/graph_dcp.html.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Hazelton, M. L. (2001). Inference for origin-destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7):667 – 676.

Maher, M. J. (1983). Inferences on trip matrices from observations on link volumes: A bayesian statistical approach. *Transportation Research Part B: Methodological*, 17(6):435–447.

Mardani, M. and Giannakis, G. (2013). Robust network traffic estimation via sparsity and low rank. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4529–4533.

Martínez, L. M., Viegas, J. M., and Silva, E. A. (2009). A traffic analysis zone definition: a new methodology and algorithm. *Transportation*, 36(5):581–599.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631.

Ortuzar, J. D. and Willumnsen, L. G. (2011). *Modeling Transport*. John Wiley and Sons, New York, 4th edition.

Sanandaji, B. M. and Varaiya, P. P. (2014). Compressive origin-destination matrix estimation. *CoRR*, abs/1404.3263.

Sheffi, Y. (1985). *Urban transportation networks: Equilibrium analysis with mathematical programming methods*. Prentice Hall Inc., New Jersey.

Slawski, M. and Hein, M. (2012). Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization.

Spiess, H. (1987). A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 21(5):395 – 412.

Tamin, O. and Willumsen, L. (1989). Transport demand model estimation from traffic counts. *Transportation*, 16(1):3–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Van-Zuylen, H. and Willumsen, L. (1980). The most likely trip matrix estimated from traffic counts. *Transportation Research Part B*, 14(3):281–293.

Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91(433):365–377.

Wang, M. and Tang, A. (2009). Conditions for a unique non-negative solution to an underdetermined system. In *47th Annual Allerton Conference on Communication, Control, and Computing, Allerton*, pages 301–307.

Wang, M., Xu, W., and Tang, A. (2011). A unique "nonnegative" solution to an underdetermined system: From vectors to matrices. *Signal Processing, IEEE Transactions on*, 59(3):1007–1016.

Willumsen, L. (1981). Simplified transport models based on traffic counts. *Transportation*, 10(3):257–278.

Yang, H., Sasaki, T., Iida, Y., and Asakura, Y. (1992). Estimation of origin-destination matrices from link traffic counts on congested networks. *Transportation Research Part B: Methodological*, 26(6):417–434.

Zhang, Y., Roughan, M., Duffield, N., and Greenberg, A. (2003). Fast accurate computation of large-scale IP traffic matrices from link loads. In *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '03, pages 206–217, New York, NY, USA. ACM.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

TABLE 1: Held-out flow prediction results of various zoning schemes and learners. The reported numbers are the mean and standard deviation over 5 independent trials. Lower values of the NRMSE and NMAE are better; higher values of ρ are better.

Zoning	**Learner**	Ω	**NRMSE**	**NMAE**	ρ
Coarse	None	NA	2.3096 ± 0.1875	2.1436 ± 0.1914	0.6532 ± 0.0338
Coarse	GLS	None	2.3497 ± 1.1177	1.7862 ± 0.7042	0.2913 ± 0.1137
Coarse	GLS	ℓ_2	0.7636 ± 0.0463	0.6479 ± 0.0581	0.7228 ± 0.0249
Coarse	GLS	ℓ_1	0.7860 ± 0.0453	0.6733 ± 0.0231	0.6898 ± 0.0317
Coarse	GLS	ℓ_1, ℓ_2	0.7653 ± 0.0208	0.6506 ± 0.0290	0.7149 ± 0.0152
Coarse	NN-GLS	None	1.2243 ± 0.4844	0.8605 ± 0.1530	0.5116 ± 0.1070
Coarse	NN-GLS	ℓ_2	0.7775 ± 0.0691	0.6542 ± 0.0389	0.6938 ± 0.0551
Coarse	NN-GLS	ℓ_1	0.8113 ± 0.0541	0.6948 ± 0.0269	0.6695 ± 0.0392
Coarse	NN-GLS	ℓ_1, ℓ_2	0.7999 ± 0.0798	0.6775 ± 0.0602	0.6702 ± 0.0734
Fine	None	NA	3.7891 ± 0.6495	3.6545 ± 0.7878	0.3526 ± 0.1090
Fine	GLS	None	0.7606 ± 0.0751	0.6416 ± 0.0252	0.7390 ± 0.0610
Fine	GLS	ℓ_2	0.6687 ± 0.0361	0.5796 ± 0.0475	0.7674 ± 0.0303
Fine	GLS	ℓ_1	0.6750 ± 0.0516	0.5722 ± 0.0462	0.7777 ± 0.0378
Fine	GLS	ℓ_1, ℓ_2	0.6896 ± 0.0661	0.5814 ± 0.0710	0.7581 ± 0.0542
Fine	NN-GLS	None	0.6940 ± 0.0552	0.5867 ± 0.0134	0.7585 ± 0.0408
Fine	NN-GLS	ℓ_2	0.6542 ± 0.0380	0.5584 ± 0.0402	0.7836 ± 0.0300
Fine	NN-GLS	ℓ_1	0.6805 ± 0.0322	0.5684 ± 0.0277	0.7787 ± 0.0352
Fine	NN-GLS	ℓ_1, ℓ_2	0.6631 ± 0.0376	0.5575 ± 0.0366	0.7804 ± 0.0289