ON SOME APPROACH TO AN ESTIMATION OF CORRESPONDENCE MATRIX OF TRANSPORT NETWORK

A. Andronov[1]

[1] *Riga Technical University,*
[1] *Riga, Latvia*
Aleksandrs.Andronovs@rtu.lv

An estimation problem of correspondence matrix for a network is considered. It is supposed that intensity of output flow for each node is fixed, but intensities of flow between nodes are unknown. It is necessary to estimate them. As a criterion of the estimation efficiency, the weighted sum of residual squares is used. Practical example is considered.

Keywords: correspondence matrix, estimation, gradient method.

1. INTRODUCTION

We have n corresponding points (towns) with numbers $i = 1, 2, ..., n$. For the point i, one is known a number h_i of inhabitants (citizen) and m numerical characteristics (categorical data) $c_{i,j}, j = 1, 2, ..., m$, those are known constants. For all pairs of the points (i, l) the distance $d_{i,l}$ between them is kown as well. Additionally, we known the size of passenger departure Y_i from point i during considered time interval, that is a random variable. Our aim is to estimate correspondence size $Y_{i,l}$ for all pairs of points (i, l), precisely the size of passenger departure from the point i to the point l. The matrix of $Y_{i,l}$ is to be said *the correspondence matrix*. Let us denote an estimate of $Y_{i,l}$ by $Y^*_{i,l}$. It is requests that all $Y^*_{i,l}$ are positive $Y^*_{i,l} > 0$ for $i \neq l$, $Y^*_{i,i} = 0$ and $Y^*_{i,l} = Y_{l,i}$. As criterion of the estimation efficiency we use the weighted sum of residual squares [3]

$$R = \sum_{i=l}^{n} w_i \Big(Y_i - \sum_{i,l} Y^*_{i,l}\Big)^2, \tag{1}$$

where w_i is a weight for the point i.

Such problem was considered earlier in the literature. Besides, usually the entropy approach is used at that. But there are many received estimates which are equal to the zero, that is inaccessible.

We use the gravitation model and the regression theory. We suppose that the concrete correspondence (i, l) for $i \neq l$ is formed with respect to model

$$Y_{i,l} = \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\Big(a + \sum_{j=l}^{m} (\alpha_j (c_{i,j} + c_{l,j}) + \beta_j (c_{i,j} c_{l,j})) + V_{i,l}\Big) \tag{2}$$

where a, $\{\alpha_j\}$ and $\{\beta_j\}$ are unknown regression coefficients,
τ and θ are unknown form coefficients,
$\{V_{i,l}\}$ are independent identically distributed random variables with zero mean and unknown variance σ^2.

As a corollary of this model we get the following presentation for the size of the passenger departure from the point i:

$$Y_i = \sum_{l=1}^{n} Y_{i,l} = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\Big(a + \sum_{j=1}^{m} (\alpha_j (c_{i,j} + c_{l,j}) + \beta_j (c_{i,j} c_{l,j})) + V_{i,l}\Big). \tag{3}$$

It is convenient to represent our models in a vector-matrix form. Let: $c_{(i)} = (c_{i,1} \ \dots \ c_{i,m})$ and $g_{(i,l)} = (c_{i,1}c_{l,1} \ \dots \ c_{i,m}c_{l,m})$ be m-vector-rows, $\alpha = (\alpha_1 \ \ \alpha_2 \dots \alpha_m)^T$,
$\beta = (\beta_1 \ \ \beta_2 \ \dots \ \beta_m)^T$.
Then

$$Y_{i,l} = \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\big(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}\big),$$

$$Y_i = \sum_{\substack{l=1 \\ i \neq l}}^{n} Y_{i,l} Y_i = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\big(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}\big). \tag{4}$$

Now we must suggest an estimate $Y^*_{i,l}$. For that we need investigate the distribution and the expectation of $Y_{i,l}$.

Note that an alternative representation of our model (4) is the following:

$$Y_{i,l} = \exp(\theta \ln(h_i h_l) - \tau \ln(d_{i,l}) + a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}),$$

$$Y_i = \sum_{l=1}^{n} Y_{i,l} = \sum_{l=1}^{n} \exp(\theta \ln(h_i h_l) - \tau \ln(d_{i,l}) + a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}). \tag{5}$$

2. DISTRIBUTION ANALYSIS

We suppose that $V_{i,l}$ has normal distribution. Then $Z_{i,l} = \exp(V_{i,l})$ has the log-normal distribution [2] with characteristics

$$E(Z_{i,l}) = E(\exp(V_{i,l})) = \exp\left(\frac{\sigma^2}{2}\right), \qquad D(Z_{i,l}) = D(\exp(V_{i,l})) = \exp(\sigma^2)(\exp(\sigma^2) - 1).$$

Therefore, for $i \neq l$

$$E(Y_{i,l}) = \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta) \exp\left(\frac{\sigma^2}{2}\right) \tag{6}$$

$$D(Y_{i,l}) = \frac{(h_i h_l)^{2\theta}}{(d_{i,l})^{2\tau}} \exp(2(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta)) \exp(\sigma^2)(\exp(\sigma^2) - 1) =$$
$$= (\exp(\sigma^2) - 1)(E(Y_{i,l}))^2. \tag{7}$$

Analogous formulae have place for $\{Y_i\}$:

$$E(Y_i) = \sum_{l=1}^{n} E(Y_{(i,l)} = \exp(\frac{1}{2}\sigma^2) \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta), \qquad (8)$$

$$D(Y_i) = \exp(\sigma^2)(\exp(\sigma^2) - 1) \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^{2\theta}}{(d_{i,l})^{2\tau}} \times$$

$$\times \exp(2(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta)) = (\exp(\sigma^2) - 1)(E(Y_i))^2. \quad (9)$$

With respect to the central limit theorem we suppose that Y_i has the normal distribution. Then, optimal weights must be chose as

$$w_i = \Big(\sum_{l=1}^{n} \frac{(h_i h_l)^{2\theta}}{(d_{i,l})^{2\tau}} \exp(2(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta))\Big)^{-1}.$$

These weights contain the unknown vector of the parameters τ, θ, a, α and β. Therefore we must use an iterative procedure and successively recalculate estimates of the weights and the parameters.

3. ESTIMATION OF REGRESSION PARAMETERS

Our model also contains the following unknown parameters: a, $\{\alpha_j\}$, $\{\beta_j\}$, τ, θ, and σ^2. If we use the expectations (8) for the estimation, then one can not identify both parameters a and σ^2 simultaneously. So, let us introduce the united parameter $\bar{a} = a + \frac{1}{2}\sigma^2$. Now instead of (6) and (8) we have the model for $i \neq l$

$$E(Y_{i,l}) = \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp(\bar{a} + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta), \qquad (10)$$

$$E(Y_i) = \sum_{\substack{l=1 \\ i \neq l}}^{n} E(Y_{i,l}) = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} exp(\tilde{a} + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta). \qquad (11)$$

Further as early,

$$D(Y_{i,l}) = (\exp(\sigma^2) - 1)(E(Y_{i,l}))^2, \qquad (12)$$

$$D(Y_{i,}) = (\exp(\sigma^2) - 1)(E(Y_i))^2. \qquad (13)$$

We will use three procedures. The first procedure is the main one. It estimates the parameters $\tilde{a}$, α and β providing fixed values of the scalar parameters τ and σ. The second procedure finds estimates of the parameters τ and σ using the simple mesh method. The third procedure identifies a and σ^2. The main procedure supposes three steps. The zero step is an initial one, the other steps are repeated.

The zero step We chose initial values of parameters $\tilde{a}$, α, β, τ, σ and calculate values of the weights $\{w_i\}$. Let $W = diag(w_1, ..., w_n)$ be a diagonal matrix with the values $\{w_i\}$ on the main diagonal.

The first step Using a gradient method [1] to find an estimates of $\tilde{a}$, α and β, which minimize the criterion (1) for fixed values τ, θ and $\{w_i\}$. We have the following expressions for the gradient of the sum (1) with respect to the scalar parameter $\tilde{a}$ and vectors of the parameters α and β:

$$\nabla R\left(\begin{pmatrix}\tilde{a}\\ \alpha\\ \beta\end{pmatrix}\right) = \begin{pmatrix}\frac{\partial}{\partial\tilde{a}}R\\ \frac{\partial}{\partial\alpha}R\\ \frac{\partial}{\partial\beta}R\end{pmatrix} =$$

$$= -2\begin{pmatrix}\sum_{i=1}^n w_i(Y_i - Y_i^*)\sum_{l=1}^n \frac{(h_ih_l)^\theta}{d_{i,l}^\tau}\exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta)\\ \sum_{i=1}^n w_i(Y_i - Y_i^*)\sum_{l=1}^n \frac{(h_ih_l)^\theta}{d_{i,l}^\tau}\exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta)(c_{(i)} + c_{(l)})^T\\ \sum_{i=1}^n w_i(Y_i - Y_i^*)\sum_{l=1}^n \frac{(h_ih_l)^\theta}{d_{i,l}^\tau}\exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta)g_{(i,l)}^T\end{pmatrix}.$$

Now we can apply a gradient method to minimize the objective function (1). Corresponding values of the vectors give the estimates $\tilde{a}^*$, α^* and β^* for the current iteration.

The second step The weights $\{w_i\}$ are recalculated for new values $\tilde{a}^*$, α^* and β^*. If a difference between two last values of sum (1) is less then prescribed precision $\varepsilon > 0$, then the iterations are ended. In other case, the transition on the first step is performed.

The described main procedure is repeated for all pairs of the values τ and θ belonging to mesh points. The best value of the objective function determines estimates τ^*, θ^*, $\tilde{a}^*$, α^* and β^*. The third procedure is described further.

4. AN ESTIMATION OF PARAMETERS α AND σ^2

According to (10) and (11) we have the estimates for $i \neq l$:

$$E(Y_{i,l})^* = \frac{(h_ih_l)^{\theta^*}}{(d_{i,l})^{\tau^*}}\exp(\tilde{a}^* + (c_{(i)} + c_{(l)})\alpha^* + g_{(i,l)}\beta^*), \tag{14}$$

$$E(Y_i)^* = \sum_{\substack{l=1\\ i\neq l}}^n E(Y_{i,l})^* = \sum_{\substack{l=1\\ i\neq l}}^n \frac{(h_ih_l)^{\theta*}}{(d_{i,l})^{\tau*}}\exp(\tilde{a}^* + (c_{(i)} + c_{(l)})\alpha^* + g_{(i,l)}\beta^*). \tag{15}$$

Analogously from (12) and (13) we get

$$D(Y_{i,l})^* = (\exp(\sigma^{2*}) - 1)(E(Y_{i,l})^*)^2, \tag{16}$$

$$D(Y_i)^* = (\exp(\sigma^{2*}) - 1)\sum_{l=1}^n (E(Y_{i,l})^*)^2. \tag{17}$$

At other hand we can estimate the variance of Y_i with respect to the variance definition

$$D(Y_i) = E(Y_i - E(Y_i))^2.$$

Using $E(Y_i)^*$ as the estimate $E(Y_i)$, we have an alternative estimate of $D(Y_i)$:

$$D(Y_i)^{**} = (Y_i - E(Y_i)^*)^2. \tag{18}$$

(Here we suppose a weak dependence between Y_i and $E(Y_i)^*$ because the last is calculated on the base of many $\{Y_i\}$).

Now the variance parameter σ^2 can be estimated using the equalization of both values (17) and (18). By summing ones for $i = 1, ..., n$, we get

$$\sum_{i=1}^{n} D(Y_i)^* = \sum_{i=1}^{n} D(Y_i)^{**} \quad \text{or} \quad \left(\exp\left(\sigma^{2*}\right) - 1\right) \sum_{i=1}^{n} \sum_{\substack{l=1 \\ l \neq i}}^{n} (E(Y_{i,l})^*)^2 = \sum_{i=1}^{n} (Y_i - E(Y_i)^*)^2.$$

Therefore

$$\sigma^{2*} = \ln \left\{ 1 + \left(2 \sum_{i=1}^{n-1} \sum_{l=i+2}^{n} (E(Y_{i,l})^*)^2\right)^{-1} \sum_{i=1}^{n} (Y_i - E(Y_i)^*)^2 \right\} \tag{19}$$

Now the estimate of the parameter a is calculated as $a^* = \tilde{a}^* - \frac{\sigma^{2*}}{2}$.

5. BALANCING

Often one requests that the statistical data $\{Y_i\}$ and the estimates $\{Y_{i,l}^*\}$ should be balanced:

$$\sum_{l=1}^{n} Y_{i,l}^* = Y_i, \quad i = 1, ..., n. \tag{20}$$

For that we introduce the corrected coefficient $\delta_i > 0$ for each point i. Then corrected estimates are

$$\tilde{Y}_{i,l} = \delta_i Y_{i,l}^* \delta_l, \quad i, l = 1, ..., n. \tag{21}$$

To calculate the coefficients $\{\delta_i\}$ we have nonlinear system

$$\delta_i \sum_{l=1}^{n} Y_{i,l}^* \delta_l = Y_i, \quad i = 1, ..., n, \quad \delta_i = \left(\sum_{l=1}^{n} Y_{i,l}^* \delta_l \right)^{-1} Y_i, \quad i = 1, ..., n.$$

The experience shows that a solution is determined simply by the method of successive approaches: for the k-th iteration

$$\delta_i^{(k)} = \left(\sum_{l=1}^{i-1} Y_{i,l}^* \delta_l^{(k)} + \sum_{l=i+1}^{n} Y_{i,l}^* \delta_l^{(k-1)} \right)^{-1} Y_i, \quad i = 1, ..., n, \tag{22}$$

where $\{\delta_i^{(0)}\}$ are initial values and we keep in mind that $Y_{i,i}^* = 0$. The iterations have ended, when a difference between two last values of $\delta^{(k)} = \left(\delta_1^{(k)} \delta_2^{(k)} \ldots \delta_n^{(k)}\right)$ is less then the prescribed precision $\varepsilon > 0$.

6. EXAMPLE

Our example concerns seven ($n = 7$) largest towns of Latvia: 1. Riga, 2. Daugavpils, 3. Jelgava, 4. Jurmala, 5. Liepaja, 6. Rezekne, 7. Ventspils. The inhabitants (citizen) numbers (in then thousand peoples) are represented by vector h:

$$h = (h_1 \ h_2 \ \dots \ h_n)^T = (76.6 \ 11.6 \ 6.4 \ 5.6 \ 9.0 \ 3.9 \ 4.4)^T.$$

As numerical characteristics (categorical data) of the i-th town, two indices have been chosen: $c_{i,1}$ - significance as a rail junction and $c_{i,2}$ - significance as a seaport. The corresponding numerical values are presented by 7×2 matrix:

$$c = \begin{pmatrix} 1 & 2 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 5 & 0 & 2 \end{pmatrix}^T.$$

The following symmetric matrix describes distances between towns (in ten km):

$$d = \begin{pmatrix} 0 & 22.9 & 4.7 & 2.3 & 21.5 & 24.2 & 18.4 \\ 22.9 & 0 & 26.6 & 26.0 & 44.4 & 9.3 & 42.1 \\ 4.7 & 26.6 & 0 & 5.7 & 17.5 & 27.9 & 17.5 \\ 2.3 & 26.0 & 5.7 & 0 & 20.1 & 27.3 & 16.6 \\ 21.5 & 44.4 & 17.5 & 20.1 & 0 & 45.7 & 11.9 \\ 24.2 & 9.3 & 27.9 & 27.3 & 45.7 & 0 & 43.4 \\ 18.4 & 42.1 & 17.5 & 16.6 & 11.9 & 43.4 & 0 \end{pmatrix}.$$

Factual passenger departure (in billion passengers) in some year is the following:

$$Y = (Y_1 \ Y_2 \ \dots \ Y_n)^T = (15.0 \ 2.46 \ 0.16 \ 8.40 \ 4.87 \ 0.37 \ 1.45)^T.$$

The above described estimation procedure gives the following estimates for $\forall w_i = 1$:

$$\theta^* = 0.706, \tau^* = 1.467, \begin{pmatrix} \bar{a}^* \\ \alpha_1^* \\ \alpha_2^* \end{pmatrix} = \begin{pmatrix} -1.42 \\ 0.229 \\ 0.169 \end{pmatrix}, \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} = \begin{pmatrix} 0.157 \\ 0.168 \end{pmatrix}.$$

Now we can calculate estimate (15):

$$Y^* = E(Y)^* = (E(Y_1)^* \ \dots \ E(Y_7)^*)^T = (14.6 \ 1.57 \ 3.47 \ 8.14 \ 4.05 \ 0.73 \ 2.25)^T.$$

It gives the value of the criteria (1) $R = 1.917$. Now we must estimate the parameters a and σ^2. Firstly, the estimate σ^{2*} is calculated by formula (19). We get $\sigma^{2*} = 0.158$. Secondly, using the estimate $\bar{a}^* = -1.42$ we find

$$a^* = \bar{a}^* - \frac{1}{2}\sigma^{2*} = -1.42 - \frac{1}{2}0.158 = -1.499.$$

Let us consider the balancing procedure (23). As initial values of coefficients we take the unit vector $\delta^{(0)} = (\delta_1^{(0)} \ \delta_2^{(0)} \ \dots \ \delta_n^{(0)}) = (1 \ 1 \ \dots \ 1)$. After 40 iterations we get values $\delta^{(40)} = (1.664 \ 1.224 \ 0.030 \ 0.679 \ 0.983 \ 0.354 \ 0.524)$, and they are not change more. Now the corrected estimates $\tilde{Y}_{i,l} = \delta_i Y_{i,l}^* \delta_l$ satisfy balance condition (20).

7. CONCLUSIONS

The problem of the estimation of correspondence size $Y_{i,l}$ for all pairs of network points (i, l) is considered. As initial data, the size of departure Y_i from point i during the considered time interval is used. The nonlinear regression model for the concrete correspondence $Y_{i,l}$ has been suggested. The unknown model's parameters have been estimated by the gradient method. Numerical examples show that considered approach gets good results.

REFERENCES

1. *Sleeper A.* Six Sigma Distribution Modeling. // McGraw Hill, New York. 2007.
2. *Srivastava M. S.* Methods of Multivariate Statistics. // Second Edition. John Wiley and Sons Inc, New York. 2002.
3. *Nocedal J, Wright S. J.* Numerical Optimization. // Second Edition. Springer, New York. 2006.