基于 GTM-TT 算法的城市区域交通状态分析

赵志强,张 毅,胡坚明

(清华大学 自动化系,北京 100084)

摘 要:将 GTM-TT 算法应用到城市区域交通状态的分析研究中,通过对北京市实际道路 29 个线圈 89 天的占有率数据进行分析,实现了高维数据的可视化和无监督聚类。在隐平面中,不同区域代表不同交通拥堵状况。统计发现频繁发生的状态转移跨度小,符合交通状态缓慢变化的认识。最后将 89 天按隐平面上的状态序列向量之间的距离进行层次聚类,聚为 4 类,分析发现了每类所代表的典型交通过程。

关键词:交通运输工程;多元时间序列;生成式拓扑映射的时间扩展;交通状态分析;无监督聚类;数据可视化

中图分类号:U491.1.12 **文献标识码**:A **文章编号**:1671-5497(2009)Sup.2-0001-06

Multi-dimensional regional traffic status analysis based on GTM-TT

ZHAO Zhi-qiang, ZHANG Yi, HU Jian-ming

(*Department of Automation, Tsinghua University, Beijing* 100084, *China*)

Abstract: The occupancy data of a small region contains 29 loops in Beijing city was analyzed using generative topographic mapping through time (GTM-TT). The original data was mapped from 29-dimensional space to 2-dimensional space which is easy for visualization and clustered into different classes naturally. The statistics of states transfer show that the most probably transfer distance was within 1 step. At last the 89 days was clustered into 4 types using hierarchical clustering method according to the Euclidean distance between the tracks of different days in the latent space. It was found that the day of different type had different traffic process.

Key words: engineering of communication and transportation; multivariate time series; GTM-TT; traffic status analysis; unsupervised clustering; data visualization

区域交通状态常采用区域内所有交通检测参数组成交通状态观测向量 t 来表征。一个普通的十字路口,如果只考虑直行方向则至少包括 4 个方向交通流;若考虑左转交通流,则检测参数会更多。一个包含了十几个路口的城市区域,其 t 往往就是一个数十维的向量。在高维空间中,难以对数据进行观察分析,所以需要一种方法将高维的数据映射到低维的空间(为便于可视化,通常为 2 维),降维的同时又能尽可能保留数据在高维空间的相对关系。并且由于区域内各交通流之间具

收稿日期:2009-05-25.

基金项目:"973"国家重点基础研究发展规划项目(2006CB705506);国家自然科学基金项目(50708054,60774034);"863"国家高技术研究发展研究计划项目(2007AA11Z222);"十一五"国家科技支撑项目(2006BAJ18B02).

作者简介:赵志强(1981-),男,博士生研究生. 研究方向:智能交通系统. E-mail:zhaozq03@tsinghua.edu.cn

通信作者:胡坚明(1975-),男,副教授,博士. 研究方向:智能交通系统. E-mail:hujm@mail.tsinghua.edu.cn

有较强的相关性，变化趋势相似，因此在高维状态空间的数据本质上是分布在一个低维流形，或说是低维隐变量空间上，交通数据压缩的研究结果[1-2]证明了这一点。所以对城市区域交通数据进行降维既是必须的，也是可行的。

目前最常用的数据降维方法是主成分分析，其通过线性旋转变换寻找方差最大的方向作为坐标轴方向，舍弃方差较小的维度，实现降维。但有时被抛弃的未必都是不重要的信息；并且 PCA 不能保证数据在低维空间相对拓扑关系不变；此外交通数据内部蕴含复杂非线性关系。因此不宜采用线性变换降维方法对交通数据进行研究。陈煜东等[3-4]使用自组织映射（Self-organizing mapping，SOM）对城市区域交通状态进行了可视化和降维，实现了向量的非线性映射和聚类，并在映射后的 2 维平面上进行交通数据的预测，取得了不错的效果。但 SOM 认为相邻输入向量彼此无关，并未考虑到交通状态是一个时间序列，相邻时刻的数据之间具有较大的相关性，所以采用 SOM 处理此问题，丢失了很多信息。

生成式拓扑映射的时间扩展（Generative topographic mapping through time，GTM-TT）算法结合了 GTM 算法与隐马尔可夫模型（Hidden markov model，HMM）。它能实现高维时间序列的降维、无监督聚类、可视化和状态序列估计，并且非线性映射到低维空间上的结果能尽可能保持数据在高维空间上的拓扑相对关系。由于区域交通状态向量内部具有的非线性和相关性，因此适宜采用 GTM-TT 算法对其进行分析研究。

1 GTM-TT 算法介绍

1.1 GTM 算法介绍

GTM 算法由 Bishop 等[5]在 1998 年提出，它近似实现了 SOM 的功能。但相比 SOM，GTM 算法有收敛性确定、存在全局损失函数、参数自动调整等优点。

设有一组 d 维空间 $\boldsymbol{T}$ 中的数据 $\boldsymbol{t}_i=(t_1,t_2,\cdots,t_d)^{\mathrm{T}}$，$i=1,2,\cdots,N$，但数据本质上是 l 维的（$d>l$），即数据 $\boldsymbol{t}_i$ 内部并不是彼此独立不相关，而是存在着某种约束关系。则数据 $\boldsymbol{t}_i$ 构成了 d 维空间上的一个 l 维的非欧子空间。为了便于展示，选择 $d=3$，$l=2$，如图 1 所示。

图 1 左边为一个 2 维平面 $\boldsymbol{x}$。$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 为其 2 个坐标轴，通过一个非线性函数 $\boldsymbol{y}(\boldsymbol{x};\boldsymbol{W})$ 将平面

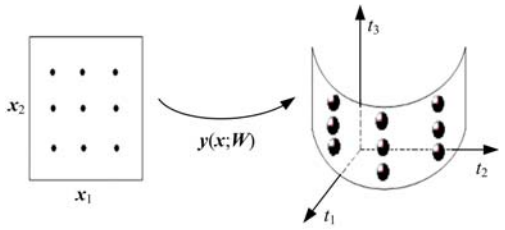

图 1 GTM 算法模型

Fig. 1 Basic idea of GTM model

映射到 3 维空间 $\boldsymbol{T}$ 中的一个低维流形曲面上，其中 $\boldsymbol{W}$ 为参数矩阵。假设隐变量 $\boldsymbol{x}$ 分布在 2 维平面 $\boldsymbol{x}$ 的 K 个点上（如图 1 左所示，K = 9），且其先验分布 $p(\boldsymbol{x})$ 是已知的。在 $\boldsymbol{T}$ 上，由于数据分布仅仅是近似位于一个低维流形上，所以映射过程存在随机噪声 $\boldsymbol{\eta}$（如图 1 左所示，9 个点映射到右边的 9 个球状分布）。

$$\boldsymbol{t}=\boldsymbol{y}(\boldsymbol{x};\boldsymbol{W})+\boldsymbol{\eta}$$

因为广义线性回归模型具有良好的逼近能力，所以映射函数做如下选择

$$\boldsymbol{y}(\boldsymbol{x};\boldsymbol{W})=\boldsymbol{W}\varphi(\boldsymbol{x})$$

式中：$\varphi(\boldsymbol{x})=[\varphi_1(\boldsymbol{x}),\varphi_2(\boldsymbol{x}),\cdots,\varphi_M(\boldsymbol{x})]^{\mathrm{T}}$。

$\varphi_m(\boldsymbol{x})$ 为 M 个中心固定且半径相同的径向基函数（Radial basis function）。$\boldsymbol{W}$ 为一个 $d\times M$ 矩阵。此模型的逼近能力与多层自适应网络相同。

采用中心位于 $\boldsymbol{y}(\boldsymbol{x};\boldsymbol{W})$ 且方差为 β^{-1} 的放射状对称高斯函数作为 $\boldsymbol{x}$ 在 d 维空间映射点 $\boldsymbol{t}$ 的概率分布，如下所示

$$p(\boldsymbol{t}\mid\boldsymbol{x},\boldsymbol{W},\beta)=\left(\frac{\beta}{2\pi}\right)^{\frac{d}{2}}\exp\left[-\frac{\beta}{2}\parallel\boldsymbol{y}(\boldsymbol{x};\boldsymbol{W})-\boldsymbol{t}\parallel^2\right]$$

对 $\boldsymbol{x}$ 进行积分可得

$$p(\boldsymbol{t}\mid\boldsymbol{W},\beta)=\int p(\boldsymbol{t}\mid\boldsymbol{x},\boldsymbol{W},\beta)\,p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

为分析和聚类的方便，$\boldsymbol{x}$ 的先验分布选择为均匀分布在 K 个点 $\boldsymbol{x}_k$ 上的 δ 函数，如下所示：

$$p(\boldsymbol{x})=\frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\delta(\boldsymbol{x}-\boldsymbol{x}_k)$$

由以上两式可得：

$$p(\boldsymbol{t}\mid\boldsymbol{W},\beta)=\frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}p(\boldsymbol{t}\mid\boldsymbol{x}_k,\boldsymbol{W},\beta)$$

而 $\boldsymbol{t}$ 的分布正是在高维数据空间中的已知观测，最大化此概率可求得 $\boldsymbol{W}$ 和 β。GTM 算法可归纳为：已知隐空间中 $\boldsymbol{x}$ 的先验分布和数据空间 $\boldsymbol{T}$ 中的数据集 R_{data}，使用期望最大化（Expectation maximization，EM）算法最大化似然函数，求解参数 $\boldsymbol{W}$ 和 β，最后求解 $\boldsymbol{x}$ 的后验分布 $p(\boldsymbol{x}\mid\boldsymbol{t})$，得到

观测数据 $\boldsymbol{t}$ 对应的隐状态 $\boldsymbol{x}$ 。

1.2　GTM-TT 算法介绍

如果数据 $\boldsymbol{t}$ 取自一个时间序列，时间邻近的数据间具有较大的相关性，则采用 GTM 算法将丢失时序信息，可将 GTM 算法与隐马尔可夫模型结合。

隐马尔可夫模型刻画了一类真实状态不可直接观察（因此称隐状态），而观察向量由隐状态按一定概率分布产生，且隐状态的转移满足马尔可夫性的随机过程。隐马尔可夫过程包括了状态转移和状态显示两个随机过程，其在语音识别、行为步态识别、生物信息学等领域得到了广泛应用。

一般的马尔可夫过程可由参数 $\lambda=(\pi,\boldsymbol{A},\boldsymbol{B})$ 描述，其中 π 为系统状态的初始分布，$\boldsymbol{A}$ 为状态转移概率矩阵，$\boldsymbol{B}$ 为观测概率矩阵。GTM 模型由参数 $\lambda=(\boldsymbol{W},\beta)$ 描述，代表了隐状态和实际观测之间的一种非线性映射，可以认为是状态显示。假设隐空间上相邻状态转移概率矩阵为 $\boldsymbol{P}_{ij}$ ，模型的初始状态分布为 π，那么 GTM-TT 模型可由参数 $\lambda=(\pi,\boldsymbol{P}_{ij},\boldsymbol{W},\beta)$ 描述。图 2 为 GTM-TT 算法的模型示意图[6]，水平向右的箭头代表了隐状态的转移，竖直向下的箭头代表了隐状态的显示。

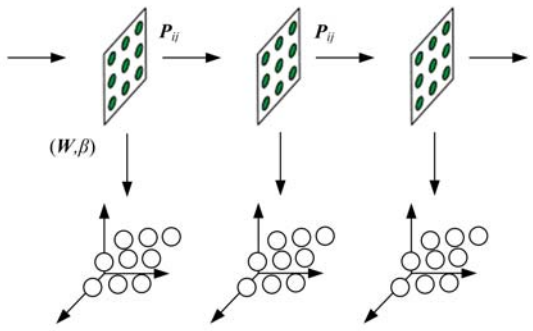

图 2　GTM-TT 算法模型

Fig. 2　Basic idea of GTM-TT model

假设实际观测数据序列为 $R_{\text{data}}=\{t_1,t_2,\cdots,t_N\}$，其对应的在潜空间内的真实状态序列为 $X_s=\{x^1,x^2,\cdots,x^N\}$，则观测到序列 R_{data} 出现的概率为

$$p(R_{\text{data}})=\sum_{x^1,x^2,\cdots,x^N}\pi_{x^1}\times\prod_{i=1}^{N}p(\boldsymbol{t}_i\mid\boldsymbol{x}^i)\times\prod_{i=1}^{N-1}p(\boldsymbol{x}^{i+1}\mid\boldsymbol{x}^i)$$

上式的计算可通过前向后向算法（forward-backward algorithm）简化。

GTM-TT 算法可归纳为：已知隐空间中 $\boldsymbol{x}$ 的先验分布和数据空间 $\boldsymbol{T}$ 中的观察序列 R_{data}，估计参数 $\boldsymbol{P}_{ij}$ 、$\boldsymbol{W}$ 和 β。GTM-TT 算法的计算过程与 GTM 算法类似，首先估测参数 $\lambda=(\pi,\boldsymbol{P}_{ij},\boldsymbol{W},\beta)$ 的初始值，然后使用 EM 算法最大化对数似然函数，求得新的模型参数值代替旧参数值，重复此过程，直至参数值收敛。这就是 Baum-Welch 算法[7]。

当 GTM-TT 模型的参数识别出来之后，就可以采用 Viterbi 算法估计最可能产生观测序列 $\{t_1,t_2,\cdots,t_N\}$ 的隐状态序列 $\{x^1,x^2,\cdots,x^N\}$ 。

2　基于 GTM-TT 算法的分析方法

2.1　研究区域介绍

研究区域选择如图 3 所示的北京市东四十条附近小区，区域内共有 30 个检测线圈，其中 12 号线圈由于故障，没有检测数据。采用剩下的 29 个线圈在时刻 i 检测的时间占有率数据组成区域状态观测向量 $\boldsymbol{t}$ ：

$$\boldsymbol{t}(i)=[Occ_1(i),Occ_2(i),\cdots,Occ_{29}(i)]^{\mathrm{T}}$$
$$i=1,2,\cdots,N$$

此处只采用了时间占有率数据，而没有采用常用的流量数据，是因为占有率在表征交通拥堵状态时，比流量更加敏感[8]。

数据采集频率为每 5 分钟 1 次，每天 288 个点。时间从 2006 年 8 月 10 日到 11 月 6 日，共 89 天。

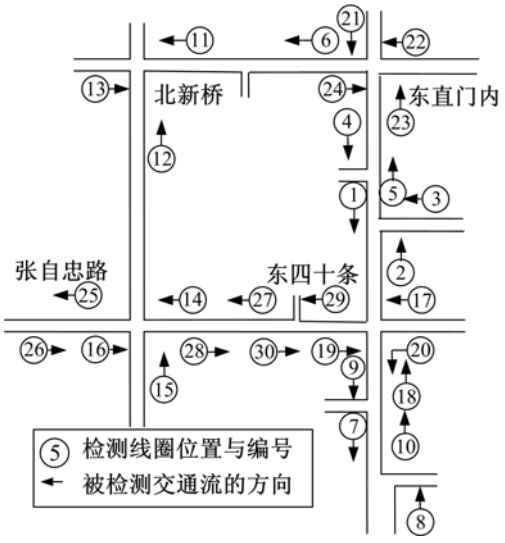

图 3　研究区域示意图

Fig. 3　Sketch map of studied traffic network region

2.2　训练过程

数据在使用之前需要进行预处理，将缺失数据和明显异常数据采用同时刻的历史平均数据替代。

在 GTM-TT 算法参数中，选择隐变量空间内格点数为 $K = 10^2$；基函数个数 $M = 4^2$；基函数的半径为其格点间距的 2 倍，从而使映射平滑，且可以保持足够的非线性映射能力。

3 分析结果与比较

3.1 无监督聚类结果

对原始观测数据使用 GTM-TT 算法训练后，可得到原 29 维空间中时间序列在 2 维隐平面[−1,1]×[−1,1]上最有可能的状态序列，也就得到了各时间点的隐状态，实现了无监督聚类，聚类个数为隐状态点数目 100。

图 4 为隐状态点在隐平面上的分布，点的灰度表示其出现时间。为了更好表示每种状态点出现的次数和时间，每个时刻的状态点坐标叠加了一个小的随机向量，使这个点随机分布在其所属的状态点方格内。从图 4 可以看到，隐平面上状态点的空间分布并不均匀，集中在有限的几点，其他点有零星分布，有些点则无分布。此外，状态点的时间分布也呈现出一定的规律性，如午夜时分，数据点基本都分布在左下角的一个方格内，而日间数据多分布在平面的中上部分。

图 5 为各状态点代表的典型向量，以及各状态点的一些统计数据。图中每一个方格代表一个隐状态点。在图中每个小方格内，曲线代表此状态点代表的典型向量，也就是 29 维占有率向量。每个方格的灰度按照占有率的平均值染色。每个方格左上角，是此状态点的编号，右上角是此状态点的出现次数，每个方格内下面的两个数字，代表了此典型向量所代表的交通状态中，占有率数值最大的两个线圈的编号，也就是最拥堵的线圈的编号。

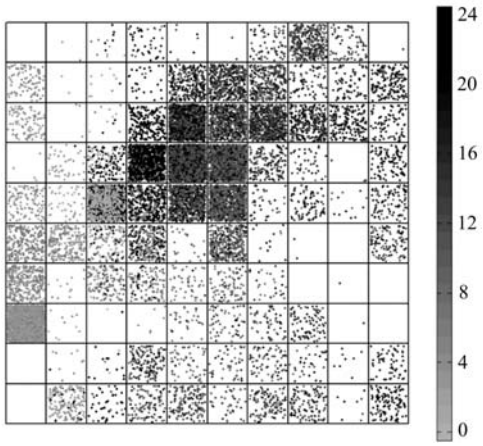

图 4 不同时间状态点在隐平面的分布

Fig. 4 Distribution of states of different time in latent space

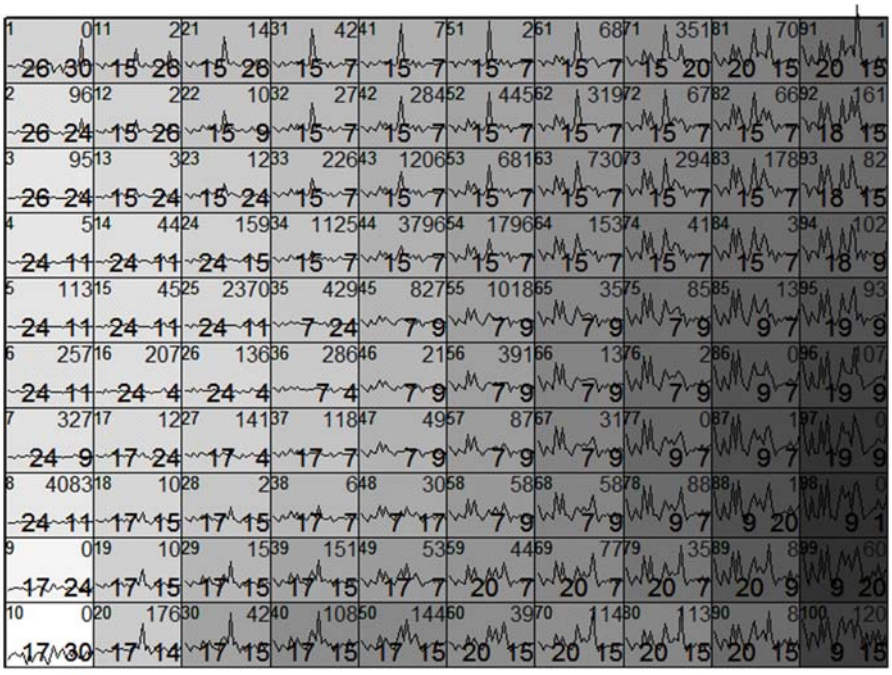

图 5 隐平面上各状态点代表的典型向量

Fig. 5 Typical vector of different states in latent space

从图 5 中可以看到，图形灰度明显左浅右深。也就是说左边代表的交通状态比较畅通，占有率低，结合图 4 可以看到，基本在午夜。而右半部分占有率高，交通拥堵，基本在日间。最右边代表特别拥堵的状态。从发生次数可以看到，交通特别拥堵的发生次数较低，通常每个点在 100 次左右，这符合交通拥堵是偶然发生的常识认识。

从图 5 中的每个状态点最拥堵的线圈序号可以看到在隐平面的上部线圈 15 都比较拥堵，尤其是在状态点 31、41、51、61 这 4 个相邻隐状态，典型状态向量是一个单峰曲线，表明只有线圈 15 处于最拥堵的状态。这 4 个状态共发生 428 次。而在隐平面的中右部，线圈 7 和 9 比较拥堵，从图 3 可知线圈 7、9 为同一道路上相连的两个路段，经过现场勘查，那里与其他 3 个方向不同，为单车道，极易发生拥堵。而在隐平面的下部，线圈 17、15、20、7、9 比较拥堵，主要是图 3 中东四十条路口向南方向交通流。由此可见，隐平面上不同的区域分别代表了不同的拥堵区域组合，不同的拥堵情况，这样就可以找到路网中的常见拥堵点组合。

3.2　动态转移分析

因为 GTM-TT 算法基于隐马尔可夫模型，所以能更好地反映状态转移的动态规律。下面对区域交通状态的动态转移特性进行分析。

图 6 为各状态转移的跨度分布，跨度按照曼哈顿距离（Manhattan distance，又称棋盘距离）衡量。从图 6 中可以看到，绝大部分的状态转移距离都小于等于 3，只有 432 次状态转移的距离大于 3，占总数的 1.69%。统计发现，发生次数最多的前 25 种状态转移只占了各种可能的转移总数的 0.25%，但其转移次数占了总共转移次数的 64.43%。频繁发生的转移模式不是停在状态点自身，就是转移向周围 1 步距离内的方格。这说明大部分的交通状态转移变化不大，符合了我们对于交通状态转移是缓慢变化的认识。另一方面，正是那些发生次数较少且跨度较大的转移，对应着交通状态的较大变化，值得注意。

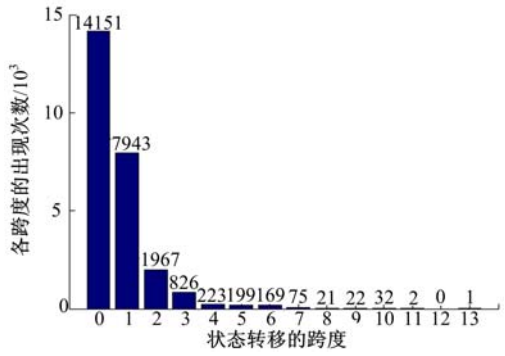

图 6　状态转移的跨度分布

Fig. 6　Distribution of state transfer distance

3.3　对不同天进行分类

每天的交通状态序列在隐平面上是一个长度为 288 的坐标轨迹序列。

$$\begin{bmatrix} x_1(1) & x_1(2) & x_1(3) & \cdots & x_1(288) \\ x_2(1) & x_2(2) & x_2(3) & \cdots & x_2(288) \end{bmatrix}$$

采用两天的时间序列之间的欧氏距离代表两天之间的距离，对 89 天采用层次聚类法进行聚类，分为 4 类。分类结果见图 7。

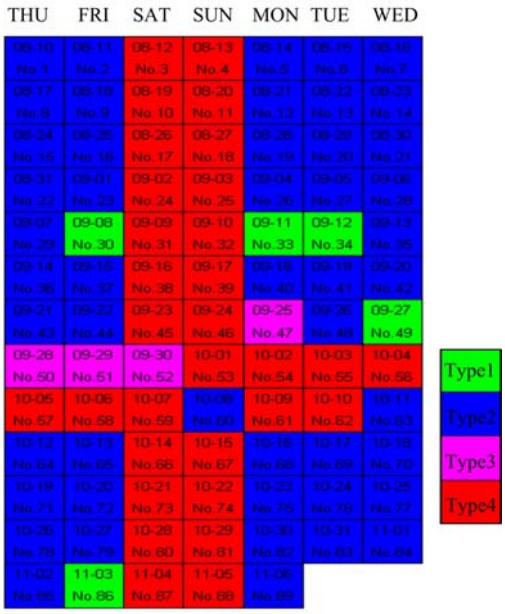

图 7　对不同天进行聚类分析的结果

Fig. 7　Clustering result of different days.

图 7 中，从类别 1 至 4，分别采用不同灰度显示。易知，类别 2 代表着平常的周中工作日，类别 4 代表着周末和节假日。

下面分别画出这 4 类天的占有率，如图 8 所示。从图 8 中可以看到，代表工作日的类别 2，其占有率明显高于其他类别，且持续时间长，尤其是与第 4 类进行比较时。这正是北京市交通流周中拥堵重，周末拥堵轻，周中与周末区别明显的表现。观察类别 1，其占有率与类别 2 相似，但是其中线圈 26～30 的占有率比其他类高，如图 8 中方框 1 所示。这 5 个线圈对应图 3 中张自忠路至东四十条区域，也就是说类别 1 代表了张自忠路至东四十条路段拥堵的工作日。观察类别 3，发现

主要是线圈 17 的占有率出现了异常波动，如图 8 中方框 2 所示。线圈 17 代表的交通流来自北京市东二环东四十条桥方向。状态 3 主要发生在国庆长假前的几天，通常是非常拥堵的。

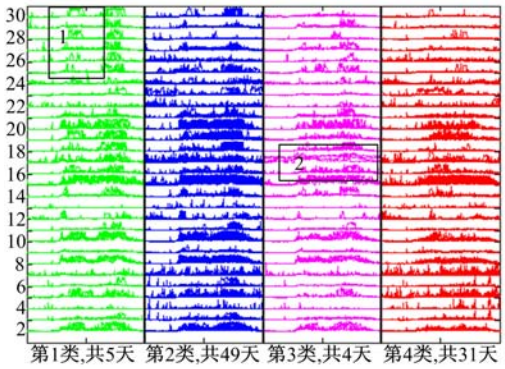

图 8 4 种不同类型的天的占有率图

Fig. 8 Occupancy of different day-type

4 结束语

使用 GTM-TT 算法对北京市区域交通状态进行了分析研究。首先得到了状态的无监督聚类结果，可将常见的不同路段拥堵组合映射到不同区域区分开来。通过对状态转移进行研究，发现发生次数最多的状态转移跨度都很短。此外，通过对隐平面上的时间序列进行聚类，将 89 天分为 4 类，每类代表了一类典型的交通拥堵状况。通过此研究，展示了采用 GTM-TT 算法对区域交通状态研究的能力。在今后的研究中，将重点研究状态动态转移的模式和机理，并将其应用到交通状态预测等方面。

致谢

作者感谢 Oxford BioSignals Ltd. Iain Guy Strachan 博士在写作过程中给与的帮助。

参考文献：

[1] Qu Li, Hu Jian-ming, Zhang Yi. A flow volumes data compression approach for traffic network based on principal component analysis[C]// Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference, Seattle, WA, USA: IEEE Omnipress, 2007: 125-130.

[2] 赵志强，张毅，胡坚明，等. 基于 PCA 和 ICA 的交通流量数据压缩方法比较研究[J]. 公路交通科技，2008，25(11)：109-115.
Zhao Zhi-qiang, Zhang Yi, Hu Jian-ming, et al. A comparison study of PCA and ICA based traffic flow compression[J]. Journal of Highway and Transportation Research and Development, 2008, 25(11): 109-115.

[3] Chen Y D, Zhang Y, Hu J, et al. Pattern discovering of regional traffic status with self-organizing maps[C]// IEEE Intelligent Transportation Systems Conference, 2006, ITSC'06, 2006: 647-652.

[4] Chen Y, Zhang Y, Hu J. Multi-dimensional traffic flow time series analysis with self-organizing maps [J]. Tsinghua Science and Technology, 2008, 13: 220-228.

[5] Bishop C M, Svensén M, Williams C K I. GTM: the generative topographic mapping [J]. Neural Computation, 1998,10: 215-234.

[6] Bishop C M, Hinton G E, Strachan I G D. GTM through time[C]// Fifth International Conference on Artificial Neural Networks(Conf. Publ. No. 440), 1997: 111-116.

[7] Olier I, Vellido A. Capturing the dynamics of multivariate time series through visualization using generative topographic mapping through time[C]// IEEE International Conference on Engineering of Intelligent Systems, 2006: 1-6.

[8] 姜桂艳，郭海峰，吴超腾. 基于感应线圈数据的城市道路交通状态判别方法[J]. 吉林大学学报：工学版，2008，38(Sup. 1)：37-42.
Jiang Gui-yan, Guo Hai-feng, Wu Chao-teng. Identification method of urban road traffic conditions based on inductive coil data[J]. Journal of Jilin University(Engineering and Technology Edtion), 2008, 38(Sup. 1): 37-42.