

隧道工作面数据挖掘分析研究

孙明宇, 王利民, 王首晨

(河北建筑工程学院, 河北 张家口 075000)

摘要: 对隧道工人时态地理信息系统 (TGIS) 数据进行研究, 可以挖掘出隧道工人在工作中潜在的信息。为解决数据噪声干扰和数据量大的困境, 通过中值滤波技术去除轨迹数据中的噪声, 使用改进的道路格拉斯算法进行数据压缩, 采用 DBSCAN 算法进行曲线聚类, 最后利用最小二乘法进行曲线拟合, 精确地找出了工作面。实验发现部分工人在一天中的轨迹具有高度相似性, 这将有助于了解工人的工作模式, 避免工人出现瞒报、误报工作情况。

关键词: 工作面; 聚类; 数据压缩; 数据挖掘; 最小二乘法

中图分类号: TP391.4; TD-05

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2025.02.014

引用格式: 孙明宇, 王利民, 王首晨. 隧道工作面数据挖掘分析研究 [J]. 网络安全与数据治理, 2025, 44(2): 88-92.

Research on data mining and analysis of tunnel working face

Sun Mingyu, Wang Limin, Wang Shouchen

(Hebei University of Architecture, Zhangjiakou 075000, China)

Abstract: Studying TGIS data of tunnel workers can uncover potential information about their work. To solve the dilemma of data noise interference and large data volume, median filtering technology was used to remove noise from trajectory data, improved Douglas algorithm was used for data compression, DBSCAN algorithm was used for curve clustering, and finally the least squares method was used for curve fitting to accurately identify the working face. Among them, it was found that some workers had highly similar trajectories in a day, which will help understand their work patterns and prevent them from concealing or misreporting their work situations.

Key words: working face; clustering; data compression; data mining; least square method

0 引言

随着土地空间信息的不断更新以及经济社会现代化的高速发展, 我国在国土综合利用、道路交通监测和治理等领域, 对地理信息系统 (Geographic Information System, GIS) 的时序分析应用有了更多的需求。时态地理信息系统 (Temporal Geographic Information System, TGIS) 已经成为 GIS 的研究热点, 它可以提供关于地理对象在时间和空间上的变化信息^[1-3]。同时, 通过对 TGIS 数据进行处理, 可以提取出不同工人的工作面, 这些数据将帮助研究人员和决策者深入了解时空现象的规律和趋势^[4], 为科学决策提供依据。

在建筑行业中, 工人是最重要的资源之一, 他们的行为直接影响项目的进度和成本。对工人的行为轨迹进行分析, 可以获取工人的工作状态, 以供管理者参考, 并相应地进行策略调查。而大多数传统的监控系统依赖

于现场工头的人工监控或者工人自行汇报^[5], 监测结果的可靠性无法保证, 且带有一定的主观性。因此, 亟需一个能自动分析的监控技术框架用来保证劳动力的高效执行。在已有研究中, 杨海军^[6]等人提出了基于 TGIS 的数字孪生智能综采工作面构建技术体系, 以直观地掌握工作面的信息; 毛善君^[7]等人基于 TGIS 的一体化等关键技术, 实现了工作面的实时传输; Zhou^[8]等人提出一个深度学习框架, 用于对多个工人的施工活动进行自动化分析。但是以上研究存在着仿真效果不佳、信息挖掘不充分等问题。

数据挖掘技术在海量数据的异常数据清洗、数据处理准确性方面具有明显的优势^[9]。如申海洋等^[10]通过提取不同时间掘进工作面的瓦斯涌出特征信息, 实现了瓦斯涌出的动态识别分析。然而, 实际应用中数据挖掘技术与 TGIS 相结合仍面临一些挑战。如处理大规模的时

空数据,如何提高数据处理的效率和性能,是一个难题;如何选择合适的方法进行处理,以保证数据的准确性和可信度,也是一个挑战^[11-12]。面对这些挑战,本文提出一种使用数据挖掘技术的方法,以某隧道工作人员为研究对象,对原始轨迹提取出关键特征,最终准确地提取出工作面。该技术框架可以排除大量冗余数据的干扰,自动分析工人每一天的工作区域,然后以工作面曲线的形式进行展示,为项目决策者提供判断依据。图1为本文的技术路线图。

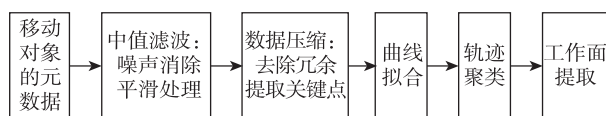


图1 技术路线图

1 工人 TGIS 数据分析与处理

1.1 TGIS 数据信息分析

本文选取某隧道勘探人员在2022年10月3日至27日的TGIS数据作为研究对象。在传感器采集的数据中,记录着不同工作人员在不同时间点所处的空间位置,对传感器采集的某一天的时空数据进行窗口分割,提取出部分TGIS数据如表1所示,其中,Time表示时间;604、755分别表示两个不同勘探人员的id;位置数据使用(x, y)的格式进行表示。

表1 原始数据样例

Time	604	755
04/10-00:00:00	(153.86, 111.12)	(-46.27, -54.56)
04/10-00:00:02	(153.91, 111.05)	(-46.27, -54.55)
.....		
04/10-22:00:04	(153.77, 112.37)	(-46.26, -54.57)
04/10-22:00:06	(153.76, 112.47)	(-46.27, -54.56)
.....		
04/10-23:00:10	(152.56, 112.66)	(-46.39, -54.51)

从表1可看出,传感器采集数据的时间间隔大概是2s,则一天的数据量就达到了4万条以上,而观察位置信息发现每个勘探人员每天不会有较大的变化。传感器在采集数据时会产生一些偏差和冗余,故在提取数据之后需要进行数据处理。

1.2 数据处理

1.2.1 中值滤波去噪

中值滤波是一个非线性数值处理方法。其原理为序列内任何一个点的取值均由该点的邻域中各点值的中值来代替^[13]。运用中值滤波进行数据处理,需要事先取得

一个窗口 m (m 为奇数),然后从输入序列中相继取出 m 个数,再将这 m 个数进行大小排序,取其序号为中心点的那个数作为滤波输出,表达式如下:

$$y_i = \text{Med} \{f_{i-v}, \dots, f_i, \dots, f_{i+v}\}, i \in N, v = \frac{m-1}{2} \quad (1)$$

其中, y_i 表示经过中值滤波后的结果, f_i 是输入序列中的第 i 个元素, v 表示窗口半径, m 表示窗口 i 的大小,Med表示中值操作,即从窗口内的所有元素中选择中值。

使用中值滤波技术处理初始矢量数据,不仅可以有效抑制孤立的噪声点,同时可以较好地保持数据的原始形状和特征,不会使矢量发生较大的变形。

1.2.2 改进的道格拉斯-普克曲线压缩

原始数据量是巨大的,即使进行了滤波技术处理,数据量依旧显得庞大、冗余,由于采集时间的频繁,多数点并不是关键路径,故本文选择基于动态规划算法(Dynamic Programming, DP)优化的道格拉斯-普克曲线压缩算法^[14]对数据进行压缩,提取出关键的特征点,方便进一步的观察和处理。

运用动态规划算法对数据的压缩优化过程如下:首先将原矢量点根据曲线压缩后的特征点进行分段线性拟合,然后通过两个矢量特征点到特征曲线的欧式距离进行误差对比,选择误差较小的曲线作为优化后的最终曲线,直到将所有曲线对比完成,并将所有曲线进行连接,得出最后的优化曲线,通过局部最优达成全局最优。因动态规划算法在进行压缩时没有对阈值进行设定,可能会导致被压缩的点距离压缩曲线的误差过大,影响整体的压缩误差,因此,在运用动态规划算法进行压缩时增加了阈值限制。动态规划压缩的公式如式(2)所示:

$$DP[n] = DP[n-1] + \min\{e(n), E(n)\} \quad (2)$$

其中, $e(n)$ 为第 n 条曲线中的矢量点到动态规划算法所求的压缩曲线的误差, $E(n)$ 为第 n 条曲线中的矢量点到逐点后退法所求的压缩曲线的误差,两者相比,取最小值,最终所得的 $DP[n]$ 即为最小压缩误差。

详细的改进压缩算法步骤如下:

(1) 对路线运用道格拉斯-普克曲线压缩算法进行压缩,得到所有的压缩线段;

(2) 以自后向前的方式运用动态规划算法对压缩曲线的特征点进行压缩;

(3) 运行动态压缩算法后,当压缩误差小于曲线压缩所得误差,而且其中的所有矢量点到曲线的阈值在曲线压缩算法所设置的阈值内,则用动态规划算法所压缩曲线进行替换,反之,则保存逐点后退法得到的压缩曲线。

图 2 为 id 为 755 的勘探人员在 10 月 3 日的压缩轨迹。

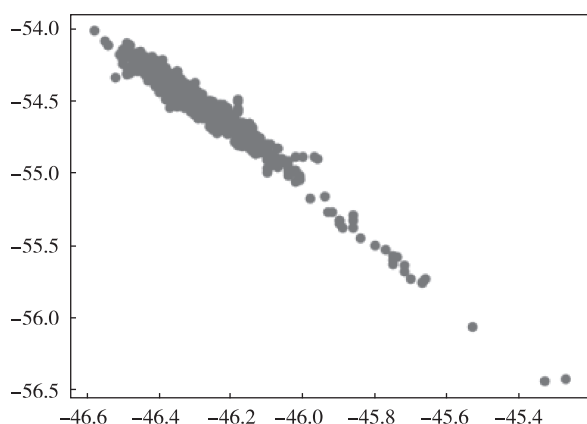


图 2 压缩轨迹

1.2.3 DBSCAN 算法聚类

本文把勘探人员一天中活动最频繁的轨迹段作为该人员当天的工作面，而活动最频繁的轨迹段密度是最大的，故选择 DBSCAN 这种基于密度的聚类算法。该算法首先会依据给出的邻域半径 Eps 和最小样本点 MinPts 进行初始化，然后对每个数据点 P ，计算其与其他数据点的距离，如果在距离范围内的数据点数量不少于最小样本点，则 P 为核心点。对于核心点，将其邻域内的点标记为同一簇，并递归地扩展该簇；对于不属于核心点但在核心点邻域内的点，标记为边界点^[15]。

DBSCAN 聚类算法中邻域半径 Eps 的取值要依据于对压缩数据绘制的 k -distance 图^[16]， k 取 2 倍数据维度减 1 (为 3) 的 k -distance 图如图 3 所示。Eps 近似取图形的拐点 0.1，而最小样本点 MinPts 取 $k+1$ 的值为 4。图 4 为该参数下 id 为 755 的勘探人员特征轨迹的聚类结果。

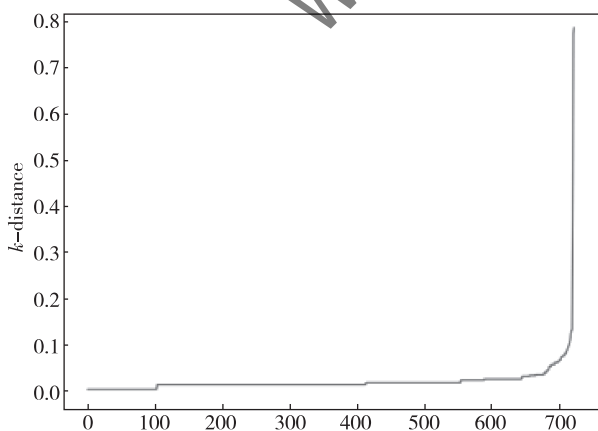


图 3 k -distance 图

聚类结果中不同密度的簇使用不同的 Cluster 进行区别。如图 4 所示，Cluster 0 为本次聚类结果中密度最大的

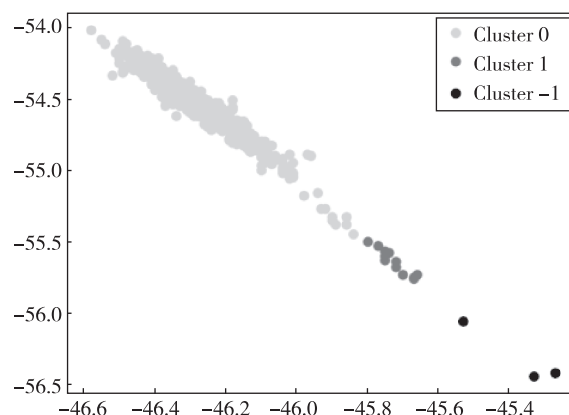


图 4 聚类结果

簇，则认为该工作人员当天的工作面区域为 Cluster 0 区域。

1.2.4 工作面提取

工作面是指一个人在工作中所面对的具体工作内容和范围，它可以代表隧道勘探工人活动轨迹的平面。而工作面在最频繁的轨迹段之中，故本文尝试拟合出一条合适的曲线代表聚类结果的原始区域。

设总体分布的位置参数为 θ ，对于给定的 α ($0 < \alpha < 1$)，如果存在两个统计量 $\theta_1(X_1, X_2, \dots, X_n)$ 与 $\theta_2(X_1, X_2, \dots, X_n)$ ，使得： $P\{\theta_1(X_1, X_2, \dots, X_n) \leq \theta \leq \theta_2(X_1, X_2, \dots, X_n)\} = 1 - \alpha$ ，则称 $[\theta_1, \theta_2]$ 为 θ 的双侧 $1 - \alpha$ 置信区间； $1 - \alpha$ 称为置信水平； θ_1 (θ_2) 称为 θ 的双侧 $1 - \alpha$ 置信区间的上 (下) 限。

设随机变量 X 的分布函数为 $F(X)$ ， p 是一个实数， $0 < p < 1$ ，若 X_p 使得：

$$P\{X \leq X_p\} = F(X_p) = p \quad (3)$$

则称 X_p 为此分布的 p 分位数。

在工作面中，取 0.95 分位数作为置信区间的上限，它使得至少有 95% 的数据项小于或等于此分位点，将 0.95 分位数作为偏移量，可以在散点图的上方使用最小二乘法绘制出一条边界线，用于表示数据的边界，并以此边界线作为其最终的工作面拟合直线。图 5 是得到的 755 号工作人员 10 月 26 日的工作面曲线。

2 实验结果与分析

2.1 工作面分析

使用该方法对不同人员不同天数的轨迹进行处理，即可得到更多的工作面，图 6、图 7 所示是提取得到的部分数据，命名方式为 id 编号_day 数字，如 639_day03 表示编号为 639 的工作人员在 2022 年 10 月 3 日的工作面。

工人每天的轨迹数据都可以用工作面来表示，以反映该工人当天主要的工作区域。一方面，工作面可以用

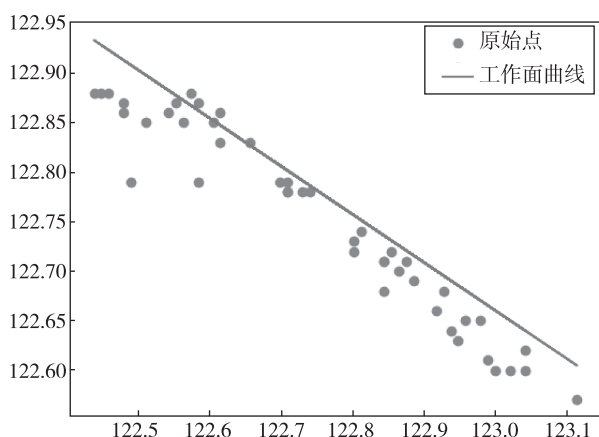


图5 工作面曲线

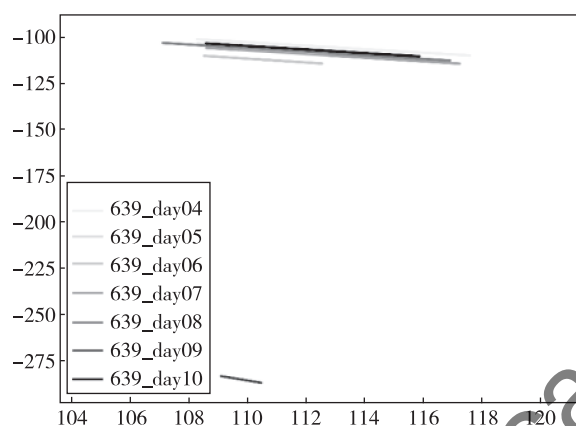


图6 639号工作人员部分工作面

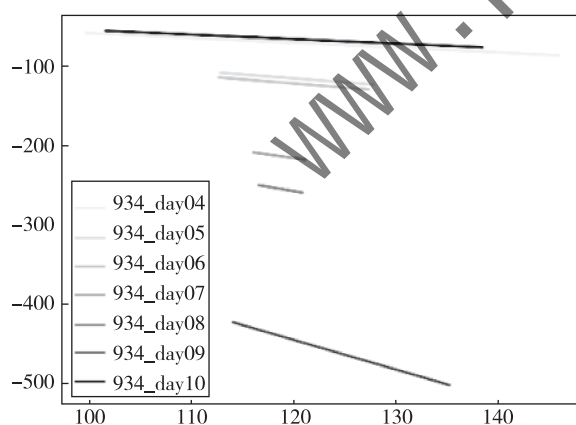


图7 934号工作人员部分工作面

来监控工人,判断某工人当天是否在指定区间工作;另一方面,工作面的不断移动或者集中在某一区间也可以用来与计划进行对比,以供管理者进行策略调查。

2.2 相似工作面分析

若某勘探人员在同一天与另一位工作人员具有类似的轨迹,即可断定他们在这一天是在同一个工作组工作

的。图8、图9所示是分析出的不同工作人员同一天类似的轨迹。

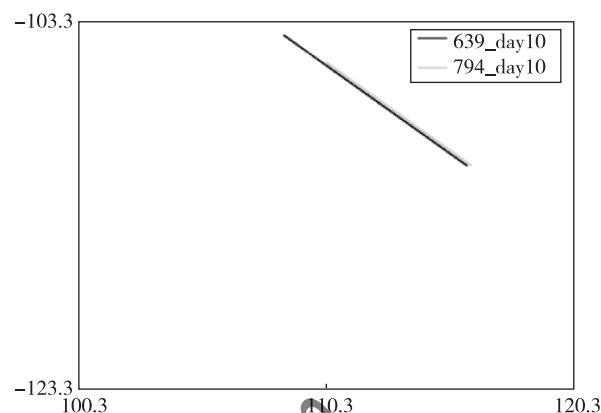


图8 10月10日部分工作面

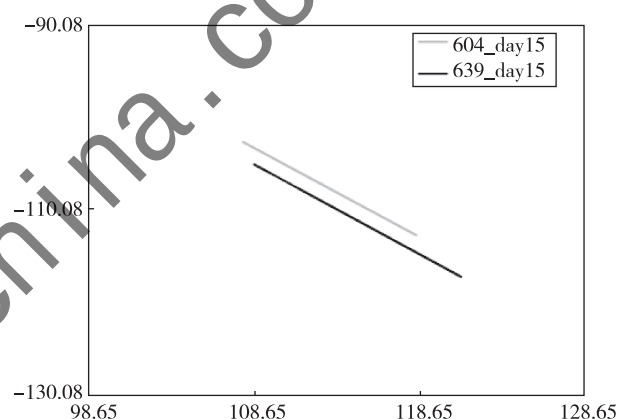


图9 10月15日部分工作面

从结果中可以看出,部分工作人员在工作中具有一些交集,这些数据可以说明部分工作面之间相互配合较好,有利于整体工程的推进。同时也给决策者一些信息提示,如资源的分配相对均衡,各个工作面能够得到所需的支持。

3 结论

本文以某隧道工作人员的工作轨迹数据为研究对象,以其工作面的提取和分析为研究目标,使用滤波技术和改进的压缩技术去除了大量轨迹数据的噪声和冗余,对关键数据进行聚类处理,准确地识别出数据中的聚类结构,最终使用最小二乘法拟合出隧道工作人员工作面的形状,并且发现部分工人轨迹存在高度相似性。通过这种方式可以自动提取工人的活动区域,避免工人汇报工作中出现瞒报、误报的情况;同时,决策者通过工人的工作面与任务计划的对比,可以对推进过快或者任务进度停滞的情况进行及时调整。

本文在提取工作面的工作中只考虑了数据的空间特

性,未来可以尝试添加如时间等更多的属性进行分析,改进压缩过程,更全面地保留压缩特征。此外,工作面曲线也可以更改为方便工作分析的其他图形,总之,对工作面的研究还需拓展领域和深度,这样才能获得更多有价值的信息。

参考文献

- [1] 毛善君,崔建军,令狐建设,等.透明化矿山管控平台的设计与关键技术[J].煤炭学报,2018,43(12):3539-3548.
- [2] 雷贵生,毛善君,王鹏飞,等.基于多维时态GIS的智慧矿区全要素云平台关键技术研究及应用[Z].延安:陕西陕煤黄陵矿业有限公司,2021-12-19.
- [3] 郭群勇,孙梅,崔磊.时空数据模型研究综述[J].地球科学进展,2016,31(10):1001-1011.
- [4] SU G S, REN H G, JIANG J Q, et al. Experimental study on the characteristics of rockburst occurring at the working face during tunnel excavation [J]. International Journal of Rock Mechanics and Mining Sciences, 2023, 164. DOI: 10.1016/j.ijrmms.2023.
- [5] YANG J, SHI Z K, WU Z Y. Vision-based action recognition of construction workers using dense trajectories [J]. Advanced Engineering Informatics, 2016, 30(3): 327-336.
- [6] 杨海军,张弘,查学刚,等.基于TGIS的数字孪生智能综采工作面研究[J].中国煤炭,2023,49(11):63-70.
- [7] 毛善君,鲁守明,李存禄,等.基于精确大地坐标的煤矿透明化智能综采工作面自适应割煤关键技术研究及系统应用[J].煤炭学报,2022,47(1):515-526.
- [8] ZHOU X H, LI S, LIU J P, et al. Construction activity analysis of workers based on human posture estimation information [J].

Engineering, 2024, 33(2): 225-236.

- [9] 乔志桃.基于出租车轨迹数据的城市居民出行时空特征挖掘与可视化分析[D].西安:长安大学,2021.
- [10] 申海洋,邹庆.大数据挖掘矿井瓦斯异常涌出系统研究[J].矿业装备,2023(11):141-143.
- [11] 蔡萌萌,张巍巍,王泓霖.大数据时代的数据挖掘综述[J].价值工程,2019,38(5):155-157.
- [12] 张友浩,赵鸣,徐梦瑶,等.时序数据挖掘的预处理研究综述[J].智能计算机与应用,2021,11(1):74-78.
- [13] 宋涛.图像中值滤波算法在智能采摘机器人中的应用[J/OL].农机化研究,1-6[2024-12-18].https://doi.org/10.13427/j.issn.1003-188X.2025.07.033.
- [14] ARI B, SOBAHI N, ALIN M F, et al. Accurate detection of autism using Douglas-Peucker algorithm, sparse coding based feature mapping and convolutional neural network techniques with EEG signals [J]. Computers in Biology and Medicine, 2022, 143: 105311.
- [15] 张学文,姚云鹏,刘江,等.基于线性DBSCAN聚类的空间群目标实时提取方法[J].空天预警研究学报,2023,37(6):422-427.
- [16] 胡洪涛.基于K-dist图的自适应参数改进的DBSCAN算法的研究与应用[D].大连:大连交通大学,2023.

(收稿日期:2024-12-18)

作者简介:

孙明宇(2001-),男,硕士研究生,主要研究方向:数据挖掘。

王利民(1974-),通信作者,男,硕士,教授,主要研究方向:数据挖掘、市政领域信息化技术。E-mail: 546507114@qq.com。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com