ACM SIGMOD PODS 2022
Philadelphia, PA, USA

AALBORG UNIVERSITET

Roskilde Universitet

MONASH University

Spatial Data Quality in the IoT Era

Management and Exploitation

Huan Li, Bo Tang, Hua Lu, Muhammad Aamir Cheema, Christian S. Jensen

4.

EXPLOITATION OF LOW-QUALITY SID

By Bo Tang

Outline

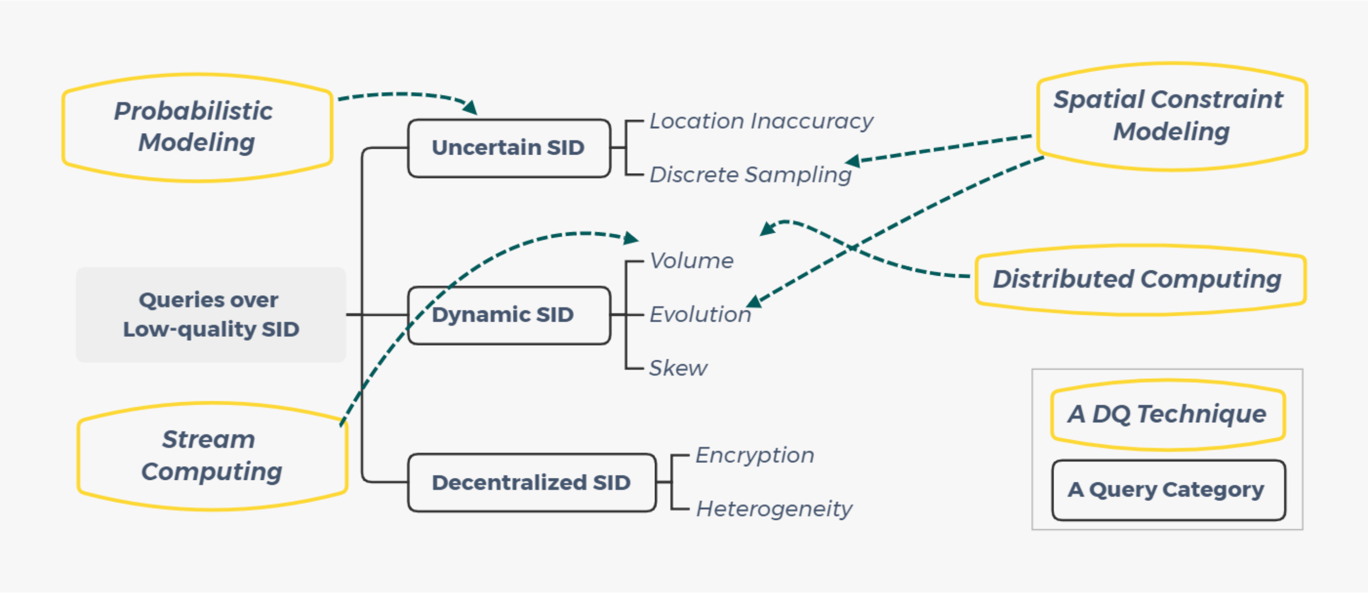
1. Queries
2. Analyses
3. Decision-making tasks

} low-quality SID

Category based on problem settings

Representatives

1. Queries over Low-Quality SID



- **Uncertainty, Dynamics, and Decentralization**

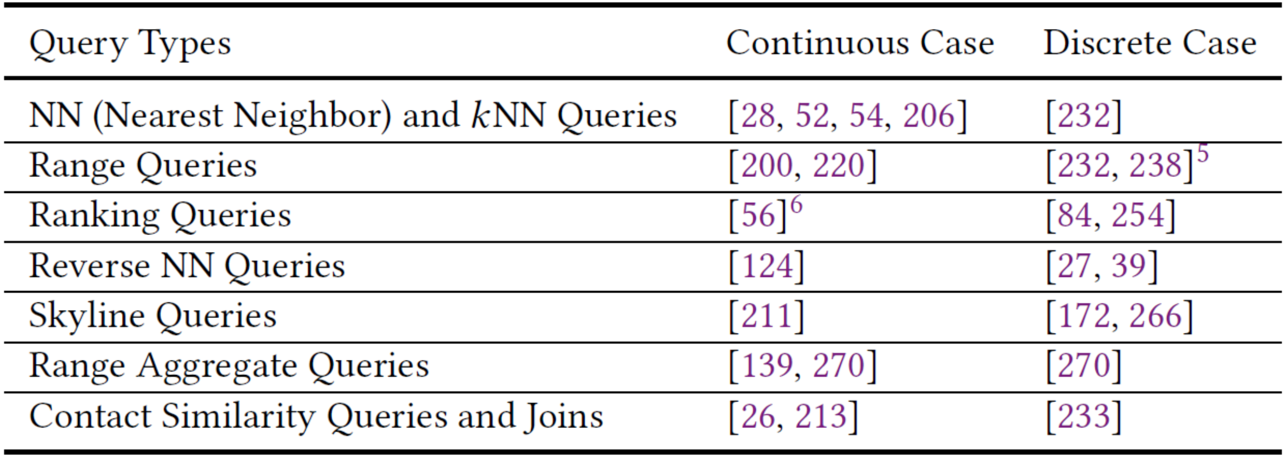
Queries over Uncertain SID

- Location Uncertainty: the major issue in *spatial queries* - Probability models [Cheng et al., 2014] [Züfle et al., 2020]
  - priority-oriented processing, object/data pruning

- Uncertainty caused by
  - **Inaccuracy of location algorithms**
  - **Discrete sampling of devices**

Queries over Uncertain SID

- ▷ Uncertainty caused by **location inaccuracy**
  - ○ a location at a time point -> probability density function
  - ○ *continuous* case: closed-form distribution
  - ○ *discrete* case: a set of samples with occurrence probabilities



Query Types	Continuous Case	Discrete Case
NN (Nearest Neighbor) and *k*NN Queries	[28, 52, 54, 206]	[232]
Range Queries	[200, 220]	[232, 238][5]
Ranking Queries	[56][6]	[84, 254]
Reverse NN Queries	[124]	[27, 39]
Skyline Queries	[211]	[172, 266]
Range Aggregate Queries	[139, 270]	[270]
Contact Similarity Queries and Joins	[26, 213]	[233]

Queries over Uncertain SID (cont.)

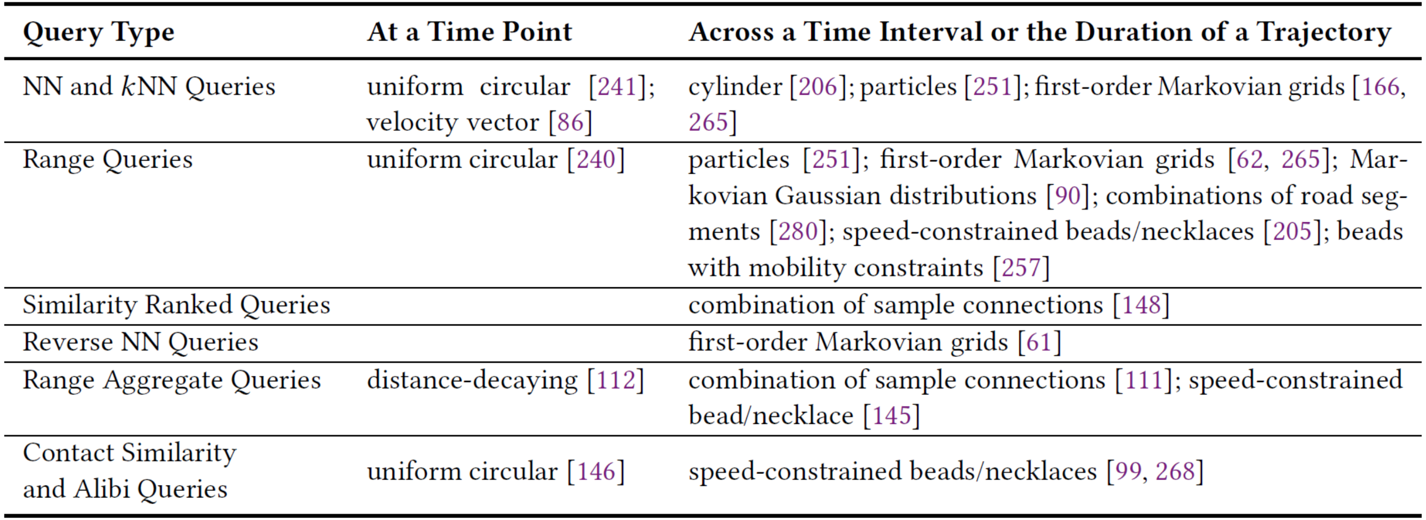
▷ Uncertainty caused by **discrete sampling**

- a location at unsampled time points -> distribution referenced to sampled, known location(s) [Pfoser et al., 1999]

infer location *at* *a single time point*? →	infer locations *across a time interval*?
• (Uniform/Gaussian/Self-defined function) circular [Yang et al., 2009] [Li et al., 2018] • Velocity vector [Huang et al., 2009]	• Particles (MCMC) [Yu et al., 2013] • First-order Markovian grids [Zhang et al., 2009] • Markovian Gaussian distribution [Jeung et al., 2013] • Combination of road segments [Zheng et al., 2011] • Beads/Necklaces [Trajcevski et al., 2010] [Kuijpers et al., 2011]

Queries over Uncertain SID (cont.)

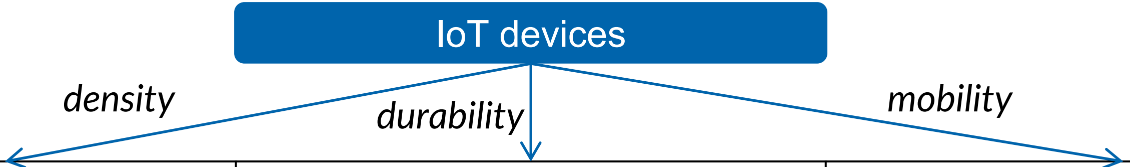
- Selected queries over uncertainty caused by **discrete sampling**



Query Type	At a Time Point	Across a Time Interval or the Duration of a Trajectory
NN and *k*NN Queries	uniform circular [241]; velocity vector [86]	cylinder [206]; particles [251]; first-order Markovian grids [166, 265]
Range Queries	uniform circular [240]	particles [251]; first-order Markovian grids [62, 265]; Markovian Gaussian distributions [90]; combinations of road segments [280]; speed-constrained beads/necklaces [205]; beads with mobility constraints [257]
Similarity Ranked Queries		combination of sample connections [148]
Reverse NN Queries		first-order Markovian grids [61]
Range Aggregate Queries	distance-decaying [112]	combination of sample connections [111]; speed-constrained bead/necklace [145]
Contact Similarity and Alibi Queries	uniform circular [146]	speed-constrained beads/necklaces [99, 268]

- **Opportunities**: resource-limited and stream setting queries?

Queries over Dynamic SID



Querying ***massive*** SID	Querying ***evolving*** SID	Querying ***skewed*** SID
• **Distributed + stream computing**: in-memory distributed similarity trajectory search [Wei et al., 2017] [Yuan et al., 2019]; Apache storm for spatio-textual queries [Mahmood et al., 2015]	• Construct a **safe region**, inside which object dynamics leave query result unchanged [Qi et al., 2018] • **Incremental evaluation** to reduce communication/computation cost at each timestamp: continuous shortest path finding [Yang et al., 2014], continuous trajectory similarity search [Zhang et al., 2020]	• **Node load-balancing**: performance heterogeneity issue [Ray et al., 2013] • **Data indexing and partitioning**: KR+-tree for skewed, multi-dimensional spatial data [Wei et al., 2014], data skew-aware partitioning with analytical strategy [Vo et al., 2014]
Opportunities: (1) high-speed yet low-cost computation? (2) decentralized edge computing nodes?		

Huan Li, Lanjing Yi, Bo Tang, Hua Lu, Christian Jensen. ***Efficient and Error-bounded Spatiotemporal Quantile Monitoring in Edge Computing Environments.*** *PVLDB 2022.*

Queries over Decentralized SID

- ▷ **Data Encryption**: encrypted outsourced data
  - ○ balance between efficiency and privacy [Yiu et al., 2010]
  - ○ dynamic data setting [Kamel et al., 2017]
  - ○ uncertain data setting [Guo et al., 2019]

- ▷ **Data Heterogeneity**: format, logics, reliability
  - ○ unified presentation for locations [Xu et al., 2013] or trajectories [Sun et al., 2017]
  - ○ unified storage and computing engine [Ding et al., 2018]

2. Analyses on Low-Quality SID

- **Uncertainty**
  - Data inaccuracy and incompleteness
- **Dynamics**
  - Volume
  - Evolution

- Clustering, Anomaly Detection, Frequent/Popular Patterns, etc.

Analyses of Uncertain SID

Task	Issue	**probabilistic modeling**	**spatiotemporal dependencies**	**constraints**
Clustering	location inaccuracy [Pelekis et al., 2011]	fuzzy vector representation		
Anomaly Events	incomplete location trace [Liu et al., 2012]	stochastic model with transition probabilities	movement as state transition	
Frequent Sequential Patterns	location inaccuracy [Li et al., 2013]	possible world	sequential explosion	
Periodic Behaviors	incomplete sequences [Li et al., 2014]	periodic behavior modeled as a probability matrix	discover reference spots	
Stop-by Patterns	noisy RFID sequences [Teng et al., 2017]	possible world	event clustering + sequential explosion	deployment and spatial constraints
Popular Routes	incomplete trajectories [Wei et al., 2012]		mutual reinforcement of collective trajectories	

▷ **Opportunities**: techniques for real-time and decentralized settings?

Analyses over Dynamic SID

- ▷ **Data Massiveness**
  - ○ **Indexing and pruning** for trajectory clustering [Wang et al., 2019], anomaly trajectory [Bu et al., 2009], co-evolving pattern mining [Zhang et al., 2015]
  - ○ **Distributed computing** for RFID trajectory clustering [Wu et al., 2014] and subsequence pattern mining [Sun et al., 2014]
- ▷ **Data Evolution**
  - ○ **Online learning** of spatiotemporal dependencies
  - ○ vehicle behavior clustering [Wang et al., 2020], anomalies in partial trajectories [Wu et al., 2017] [Liu et al., 2020]
- ▷ **Opportunities**: services to IoT edge, reduce cost/latency?

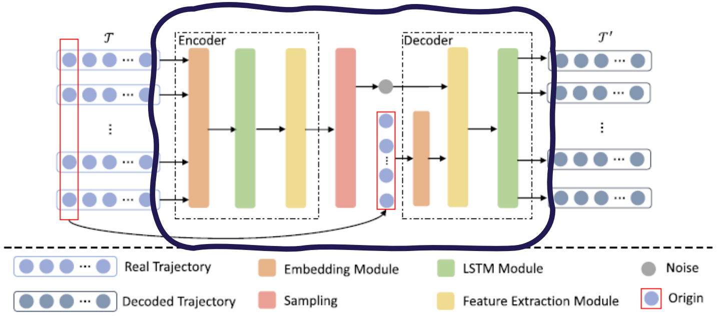
3. Decision-making using Low-Quality SID

- ▷ Predictions, Recommendations, Planning, etc.

- ▷ **Scarcity of labels**
- ▷ **Limited availability and bias of data**
- ▷ **Uncertainty of data**
- ▷ **Dynamics of data**
- ▷ **Heterogeneity and decentralization of data**

Scarcity of Labels

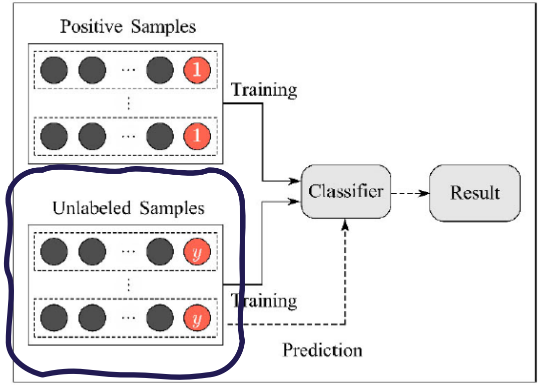
[Chen et al., 2021]



Variational AutoEncoders
for trajectory generation

No labels

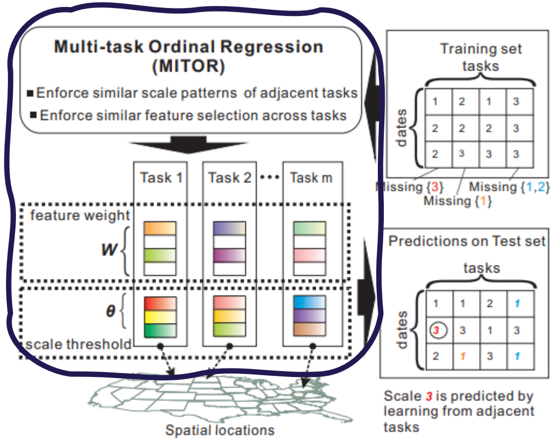
[Chen et al., 2020]



Positive-Unlabeled Learning
for site selection

Imbalanced labels

[Gao and Zhao, 2018]

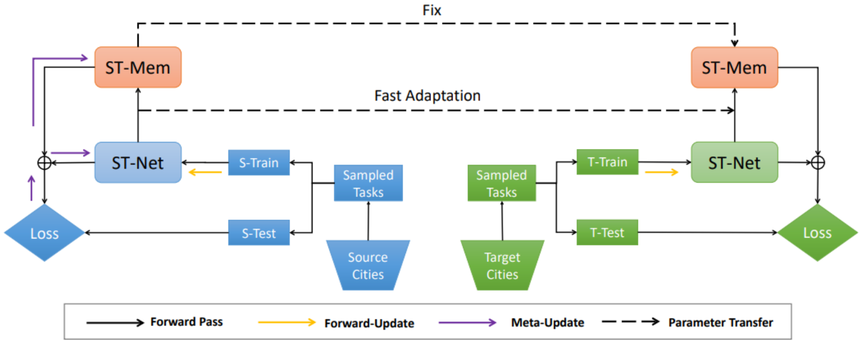

Multi-task Ordinal Regression (MITOR)
■ Enforce similar scale patterns of adjacent tasks
■ Enforce similar feature selection across tasks
Task 1
Task 2
Task m
feature weight
W
θ
scale threshold
Spatial locations
Training set tasks
dates
1 2 1 3
2 2 2 3
2 3 3 3
Missing {3}
Missing {1,2}
Missing {1}
Predictions on Test set tasks
dates
1 1 2 1
3 3 1 3
2 1 3 1
Scale 3 is predicted by learning from adjacent tasks

Multi-task Learning
for event scale prediction

Incomplete labels

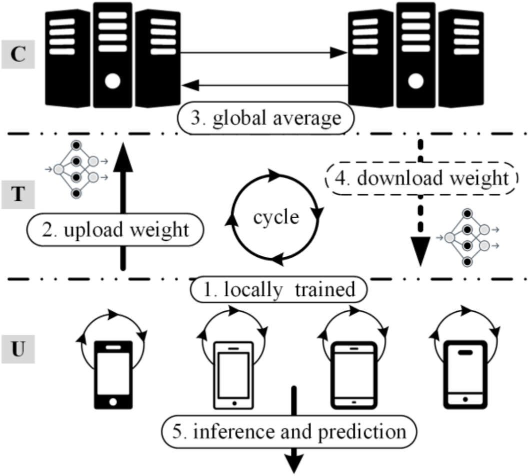
Limited Availability and Bias of Data

[Yao et al., 2019]



Meta Learning
for spatiotemporal prediction
(support multiple source cities)

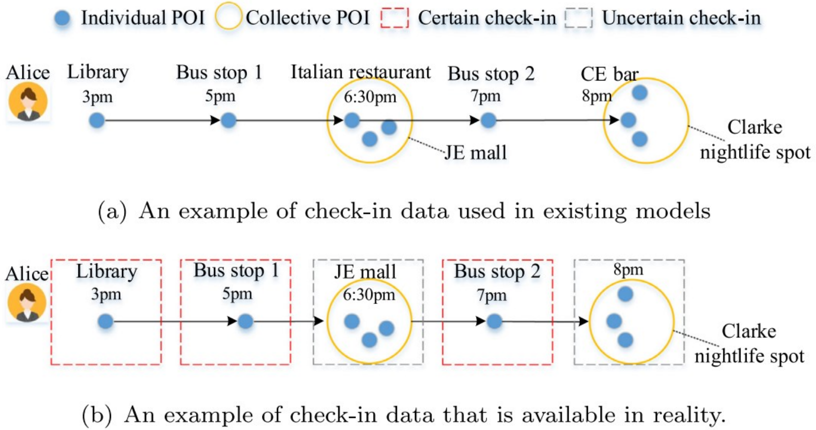
[Feng et al., 2020]



Federated Learning
for human mobility prediction
(privacy-preserving)

Uncertainty of Data

[Zhang et al., 2020-2]

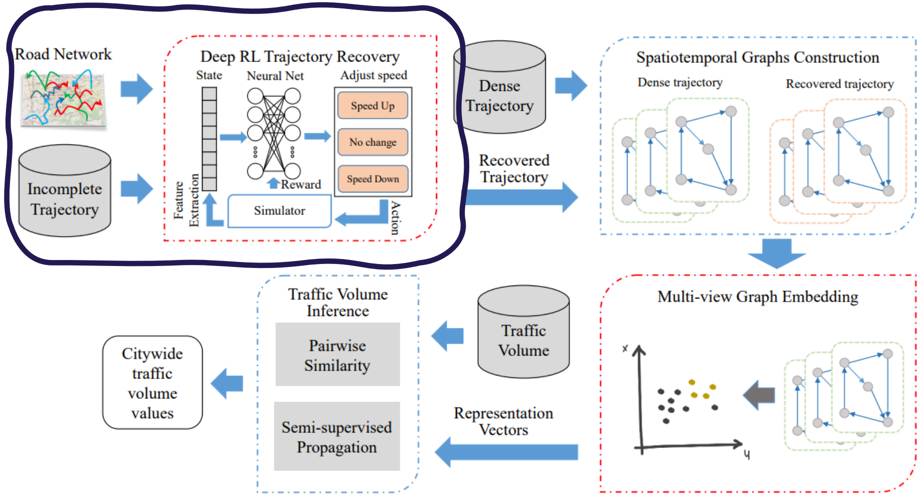


(a) An example of check-in data used in existing models

(b) An example of check-in data that is available in reality.

Probabilistic hierarchical category transitions for next PoI recommendation (coarser-grained and incomplete check-ins)

[Tang et al., 2019]



Reinforcement Learning based trajectory recovery for traffic volume prediction (incomplete radio network trajectories)

Dynamics of Data

- ▷ changing environment: accuracy
  - ○ **reinforcement learning** [Sun et al., 2021] for adaptive strategies in spatial task assignment

- ▷ streaming setting: latency
  - ○ **incremental learning** [Laha et al., 2018] updates model parameters in batches
  - ○ **edge-routing-central architecture** [Luo et al., 2019] to handle high-density IoT data

Heterogeneity and Decentralization of Data

- ▷ Fusing multi-source spatiotemporal data
  - ○ **multi-task learning** [Nguyen et al., 2019]: data aggregated at shared layers, labeled separated for different tasks
  - ○ **multi-view learning** [Zhang et al., 2015]: mutually-reinforced knowledge from multi-view data
- ▷ Building decentralized models
  - ○ **federated learning** [Liu et al., 2020-2]: secure parameter aggregation, global and local models

References

- ▷ [Cheng et al., 2014] Managing uncertainty in spatial and spatio-temporal data. *ICDE*.
- ▷ [Züfle et al., 2020] Uncertain spatial data management: An overview.
- ▷ [Pfoser et al., 1999] Capturing the uncertainty of moving-object representations. *SSD*.
- ▷ [Yang et al., 2009] Scalable continuous range monitoring of moving objects in symbolic indoor space. *CIKM*.
- ▷ [Huang et al., 2009] Continuous k-nearest neighbor query for moving objects with uncertain velocity. *GeoInformatica*.
- ▷ [Li et al., 2018] In search of indoor dense regions: An approach using indoor positioning data. *TKDE*.
- ▷ [Yu et al., 2013] An RFID and particle filter-based indoor spatial query evaluation system. *EDBT*.

References

- [Zhang et al., 2009] Effectively indexing uncertain moving objects for predictive queries. *PVLDB*.
- [Jeung et al., 2013] Managing evolving uncertainty in trajectory databases. *TKDE*.
- [Zheng et al., 2011] Probabilistic range queries for uncertain trajectories on road networks. *EDBT*.
- [Trajcevski et al., 2010] Uncertain range queries for necklaces. *MDM*.
- [Kuijpers et al., 2011] An analytic solution to the alibi query in the space - time prisms model for moving object data. *IJGIS*.
- [Xie et al., 2017] Distributed trajectory similarity search. *PVLDB*.
- [Yuan et al., 2019] Distributed in-memory trajectory similarity search and join on road network. *ICDE*.

References

- ▷ [Mahmood et al., 2015] Tornado: A distributed spatio-textual stream processing system. *PVLDB*.
- ▷ [Qi et al., 2018] Continuous spatial query processing: A survey of safe region based techniques. *CSUR*.
- ▷ [Yang et al., 2014] CANDS: Continuous optimal navigation via distributed stream processing. *PVLDB*.
- ▷ [Zhang et al., 2020] Continuous trajectory similarity search for online outlier detection. *TDKE*.
- ▷ [Ray et al., 2013] A parallel spatial data analysis infrastructure for the cloud. *SIGSPATIAL*.
- ▷ [Wei et al., 2014] Indexing spatial data in cloud data managements. *Pervasive Mob. Comput.*.

References

▷ [Vo et al., 2014] SATO: A spatial data partitioning framework for scalable query processing. *SIGSPATIAL*.

▷ [Yiu et al., 2010] Enabling search services on outsourced private spatial data. *VLDBJ*.

▷ [Kamel et al., 2017] Dynamic spatial index for efficient query processing on the cloud. *J. Cloud Comput.*.

▷ [Guo et al., 2019] Secure and efficient k nearest neighbor query over encrypted uncertain data in cloud-IoT ecosystem. *IEEE IoTJ*.

▷ [Xu et al., 2013] A generic data model for moving objects. *GeoInformatica*.

▷ [Sun et al., 2017] A real-time similarity measure model for multi-source trajectories. *CIIS*.

▷ [Ding et al., 2018] UlTraMan: A unified platform for big trajectory data management and analytics. *PVLDB*.

References

- [Pelekis et al., 2011] Clustering uncertain trajectories. *KAIS*.
- [Liu et al., 2012] A stochastic model for context-aware anomaly detection in indoor location traces. *ICDM*.
- [Li et al., 2013] Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases. *ICDM*.
- [Li et al., 2014] Mining periodicity from dynamic and incomplete spatiotemporal data. *Data Mining and Knowledge Discovery for Big Data*.
- [Teng et al., 2017] Toward mining stop-by behaviors in indoor space. *TSAS*.
- [Wei et al., 2012] Constructing popular routes from uncertain trajectories. *KDD*.
- [Wang et al., 2019] Fast large-scale trajectory clustering. *PVLDB*.
- [Bu et al., 2009] Efficient anomaly monitoring over moving object trajectory streams. *KDD*.

References

- ▷ [Zhang et al., 2015] Assembler: Efficient discovery of spatial co-evolving patterns in massive geo-sensory data. *KDD*.
- ▷ [Wei et al., 2012] Constructing popular routes from uncertain trajectories. *KDD*.
- ▷ [Wu et al., 2014] A cloud-friendly RFID trajectory clustering algorithm in uncertain environments. *TPDS*.
- ▷ [Sun et al., 2014] Mining uncertain sequence data on Hadoop platform. *PAKDD*.
- ▷ [Wang et al., 2020] Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles. *TIST*.
- ▷ [Wu et al., 2017] A fast trajectory outlier detection approach via driving behavior modeling. *CIKM*.
- ▷ [Liu et al., 2020] Online anomalous trajectory detection with deep generative sequence modeling. *ICDE*.

References

- ▷ [Chen et al., 2021] TrajVAE: A Variational AutoEncoder model for trajectory generation. *Neurocomputing*.
- ▷ [Chen et al., 2020] ToiletBuilder: A PU learning based model for selecting new public toilet locations. *IEEE IoTJ*.
- ▷ [Gao et al., 2020] Incomplete label multi-task ordinal regression for spatial event scale forecasting. *AAAI*.
- ▷ [Yao et al., 2019] Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. *WWW*.
- ▷ [Feng et al., 2020] PMF: A privacy-preserving human mobility prediction framework via federated learning. *UbiComp*.
- ▷ [Zhang et al., 2020-2] Modeling hierarchical category transition for next POI recommendation with uncertain check-ins. *Inf. Sci.*.

References

- ▷ [Tang et al., 2019] Joint modeling of dense and incomplete trajectories for citywide traffic volume inference. *WWW*.
- ▷ [Sun et al., 2021] Deep reinforcement learning for task assignment in spatial crowdsourcing and sensing. *IEEE Sens. J.*.
- ▷ [Laha et al., 2020] Real time location prediction with taxi-GPS data streams. *Transp. Res. Part C Emerg.*.
- ▷ [Luo et al., 2019] A short-term energy prediction system based on edge computing for smart city. *Future Gener. Comput. Syst.*.
- ▷ [Nguyen et al., 2019] Spatial-temporal multi-task learning for within-field cotton yield prediction. *PAKDD*.
- ▷ [Zhang et al., 2015] coMobile: Real-time human mobility modeling at urban scale using multi-view learning. *SIGSPATIAL*.

References

- ▷ [Liu et al., 2020-2] Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE IoTJ*.