Scalable Detection of Traffic Congestion from Massive Floating Car Data Streams

Győző Gidófalvi and Can Yang
Division of Geoinformatics
Deptartment of Urban Planning and Environment
KTH Royal Institution of Technology, Sweden
{gyozo,cyang}@kth.se

ROYAL INSTITUTE OF TECHNOLOGY

Outline

- Introduction
- Related work
- Method
  - Grid-based directional flow statistics
  - Directional congestion detection
  - SQL-based implementation
- Empirical evaluations
  - Quality assessment
  - Scalability assessment
- Conclusion and future work

Introduction

- Congestion is a serious problem
  - Economic losses and quality of life degradation that result from increased and unpredictable travel times
  - Increased level of carbon footprint that idling vehicles leave behind
  - Increased number of traffic accidents that are direct results of stress and fatigue of drivers that are stuck in congestion

- Road network expansion is not a sustainable solution

- Instead, utilize increasingly available Floating Car Data (FCD) to: monitor → understand → control movement and congestion

Modern Traffic Prediction and Management System (TPMS)

- Motivated by:
  - Widespread adoption of online GPS-based on-board navigation systems and location-aware mobile devices
  - Movement of an individual contains a high degree of regularity
- Use vehicle movement data as follows:
  - Vehicles periodically send their location (and speed) to TPMS
  - TPMS extracts traffic / mobility patterns from the submitted information
  - TPMS uses traffic / mobility patterns + current / recent historical locations (and speeds) of the vehicles for:
    - Short-term traffic prediction and management:
      - Predict near-future locations of vehicles and near-future traffic conditions
      - Inform the relevant vehicles in case of an (actual / predicted) event
      - Suggest how and which vehicles to re-route in case of an event
    - Long-term traffic and transport planning

Approach, Unique Features, and Contributions

- Use a data-driven approach grid-based, time-inhomogeneous model, method for the detection of congestion from large FCD streams
- Unique features
  - Grid-based model: no need to road network information and can be easily scaled to any geographical level of detail
  - Representation flow direction on the grid
  - Time-inhomogeneous
  - Novel congestion definition
  - Simple, scalable, portable SQL-based implementation

Outline

- Introduction
- Related work
- Method
  - Grid-based directional flow statistics
  - Directional congestion detection
  - SQL-based implementation
- Empirical evaluations
  - Quality assessment
  - Scalability assessment
- Conclusion and future work

Related Work: Congestion Detection

- Data sources used
  - Loop detectors, cameras, video, GPS / FCD
- Detection granularity
  - Road link, 2D region / grid cell
- Congestion metrics / indicators
  - Uniform threshold based on link travel speed
  - Ratio of average travel speed and the link's speed limit
  - Travel speed in conjunction with object density
  - Ratio of observed and expected travel time
  - Difference in travel time between two consecutive periods
- Congestion models
  - Microscopic
  - Macroscopic / pattern-based: redcurrant, clustered, dropping

Outline

- Introduction
- Related work
- Method
  - Grid-based directional flow statistics
  - Directional congestion detection
  - SQL-based implementation
- Empirical evaluations
  - Quality assessment
  - Scalability assessment
- Conclusion and future work

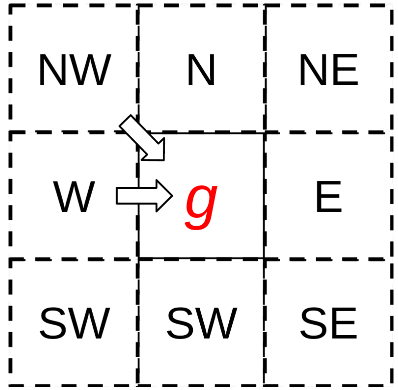
ROYAL INSTITUTE OF TECHNOLOGY

Method Outline

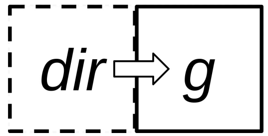
1. Map the directional flow / movement of objects to the grid-based framework.
2. Form tumbling windows over the mapped input stream and treat them as ***temporal analysis windows***.
3. Extract ***Current Directional Flow Statistics (CDFS)*** from the ***Recent Trajectories (RT)*** that are within the current tumbling / temporal analysis window.
4. Incrementally summarize the CDFS into ***Historical Directional Flow Statistics (HDFS)*** for different *temporal domain projections*.
5. Detect a grid cell *g* to be congested from a particular direction *dir* if the current mean speed of vehicles that have entered the grid cell *g* from the direction *dir* is significantly and substantially below the normal according to the temporally relevant HDFS.

Grid-based Directional Flow and Mobility Statistics

- Directional flow and movement: grid cell and its immediate 8 neighbors



- Directional flow statistics for a grid cell-direction combination (*g*, *dir*):
  - # of objects in (*g*, *dir*)
  - Average speed of objects in (*g*, *dir*)
  - Standard deviation of speeds of objects in (*g*, *dir*)



Directional Congestion Detection

- Define a grid cell-direction combination (*g*, *dir*) as a *directional congestion* based on the current $(\dot{n}, \dot{\mu}, \dot{\sigma})$ and historical $(\bar{n}, \bar{\mu}, \bar{\sigma})$ directional flow statistics if the following four criteria are satisfied:

1. Sample size criterion: $\dot{n} >= min\_veh$
2. Sample dispersion criterion: $\dot{\sigma}/\dot{\mu} < max\_cv$
3. Statistical power criterion: $(\dot{\mu} - \bar{\mu})/(\bar{\sigma}/\sqrt{\dot{n}}) < max\_z$
4. Speed difference criterion: $(\dot{\mu} - \bar{\mu})/\bar{\mu} < max\_relspddiff$

SQL: Schema

- Three database tables:

RT = <oid, dgid, spd>
CDFS = <dgid, nr, mu, sig>
HDFS = <dgid, nr, mu, sig>

- Directional grid ID **`dgid`** columns contain an integer concatenation of grid coordinates and direction (**`gx`**, **`gy`**, **`dir`**)
- Underline denotes DB indexes

SQL: Calculation of CDFS

SQL 1 FUNCTION calc_CDFS()

```
SELECT dgid, count(*) AS nr, avg(spd) AS mu,
       COALESCE(stddev(spd),0) AS sig
FROM RT
GROUP BY dgid;
```

SQL: Incremental Calculation of HDFS

SQL 2 FUNCTION ud_HDFS()

```
UPDATE HDFS AS gh
SET nr = (c.nr+gh.nr),
    mu = (c.nr*c.mu+gh.nr*gh.mu)/(c.nr + gh.nr),
    sig = sqrt((gh.nr * gh.sig^2 + c.nr * c.sig^2) /
               (gh.nr + c.nr) +
               (gh.nr * c.nr * (gh.sig - c.sig)^2) /
               (gh.nr + c.nr)^2)
FROM CDFS AS c
WHERE gh.dgid = c.dgid;

INSERT INTO HDFS (dgid, nr, mu, sig)
SELECT c.gid, c.dir, c.nr, c.mu, c.sig
FROM CDFS AS c
LEFT JOIN HDFS AS gh
ON (gh.dgid = c.dgid)
WHERE gh.dgid IS NULL;
```

No previous HDFS

- Incrementally update previously observed HDFS based on non-overlapping subset / tumbling window statistics
- Insert new / not-yet-observed statistics

SQL: Calculation of Directionally Congested Cells

SQL 3 FUNCTION CongCells(min_veh, max_cv, max_z, max_relspddiff)

```
1  SELECT c.dgid AS dgid
2  FROM HDFS AS gh, CDFS AS c
3  WHERE gh.dgid = c.dgid
4    AND c.nr >= min_veh
5    AND c.sig / c.mu < max_cv
6    AND (c.mu - gh.mu) / (gh.sig / sqrt(c.nr)) < max_z
7    AND (c.mu - gh.mu) / gh.mu < max_relspddiff;
```

Directional congestion criteria (4-7)

Temporal Domain Projections

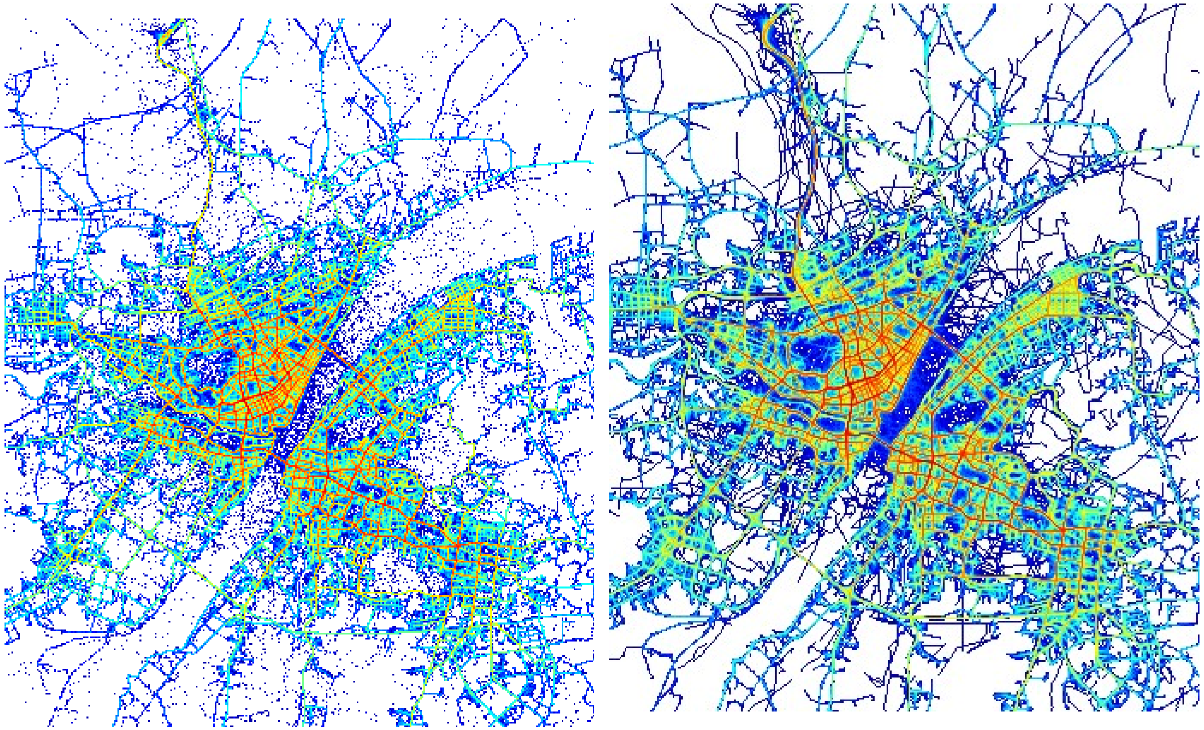
- To capture temporal regularities in flows and movements the proposed method extracts HDFS for different values of day-of-week and hour-of-day temporal domain projections
- Clients calculate **`dow`** and **`how`** projections of their status reports
- The **`HDFS`** table stores the domain projected aggregates using the value `-1` to denote the "any" value
- Detection query combines a disjunction of conditions using the relevant domain projected information in the decision criteria
  - Detection if the statistical power criterion and the speed difference criterion are satisfied either based on the **`dow`**-projected, the **`hod`**-projected or the global statistics

Outline

- Introduction
- Related work
- Method
  - Grid-based directional flow statistics
  - Directional congestion detection
  - SQL-based implementation
- Empirical evaluations
  - Quality assessment
  - Scalability assessment
- Conclusion and future work

Empirical Evaluations: Environment + Data

- Environment: 64bit Ubuntu 14.04 LTS with PostgreSQL 9.3.9 on a PC with Intel Core i7-5600U @ 2.60GHz × 4 CPU, 16GB main memory and 512GB SSD
- Data set: 6 day sample of 11K taxis in Wuhan, China (85M records)
  - Outlier removal
  - 18km x 18km city center
  - Sampling gaps of more the 120 seconds delimit trips
  - Linear interpolation of trips between samples
  - Eliminate short trips (less than 300 seconds / 10 100m-grids)
  - → 2 million trips that have an average length of 1268 seconds and 82 grid cells; ~185M status reports



Raw sample vs. interpolated trips

Empirical Evaluations: Setup

- Quality + scalability assessments
- Default parameters:
  - temporal analysis window size / prediction horizon: $\Delta t_{awin} = \Delta t_{pred}$ = 60 seconds
  - minimum number of current status reports: *min_veh* = 2
  - maximum sample dispersion: *max_cv* = 0.5
  - maximum negative z-score: *max_z* = -1.65 (significance level of α = 0.05)
  - maximum negative relative speed difference: *max_relspddiff* = -0.5
- Quality measures - traffic and congestion indicators
  - *TL*: avg. # of object present in a period (24h vs avg. temporal analysis window (TAW))
  - *NrC*: # of times a non-directional grid cell is congested in a period (24h vs avg. TAW)
  - *AbsCL* / *RelCL*: sum of absolute / relative deviation in speed from normal that objects experience in a period (24h vs avg. TAW)
- Scalability measures: time and storage (# of DB rows) that the computation phases use
  - Temporal data alignments: hod (in load exp.) vs fixed (in resolution exp.)

Quality Assessment: Spatial Distribution

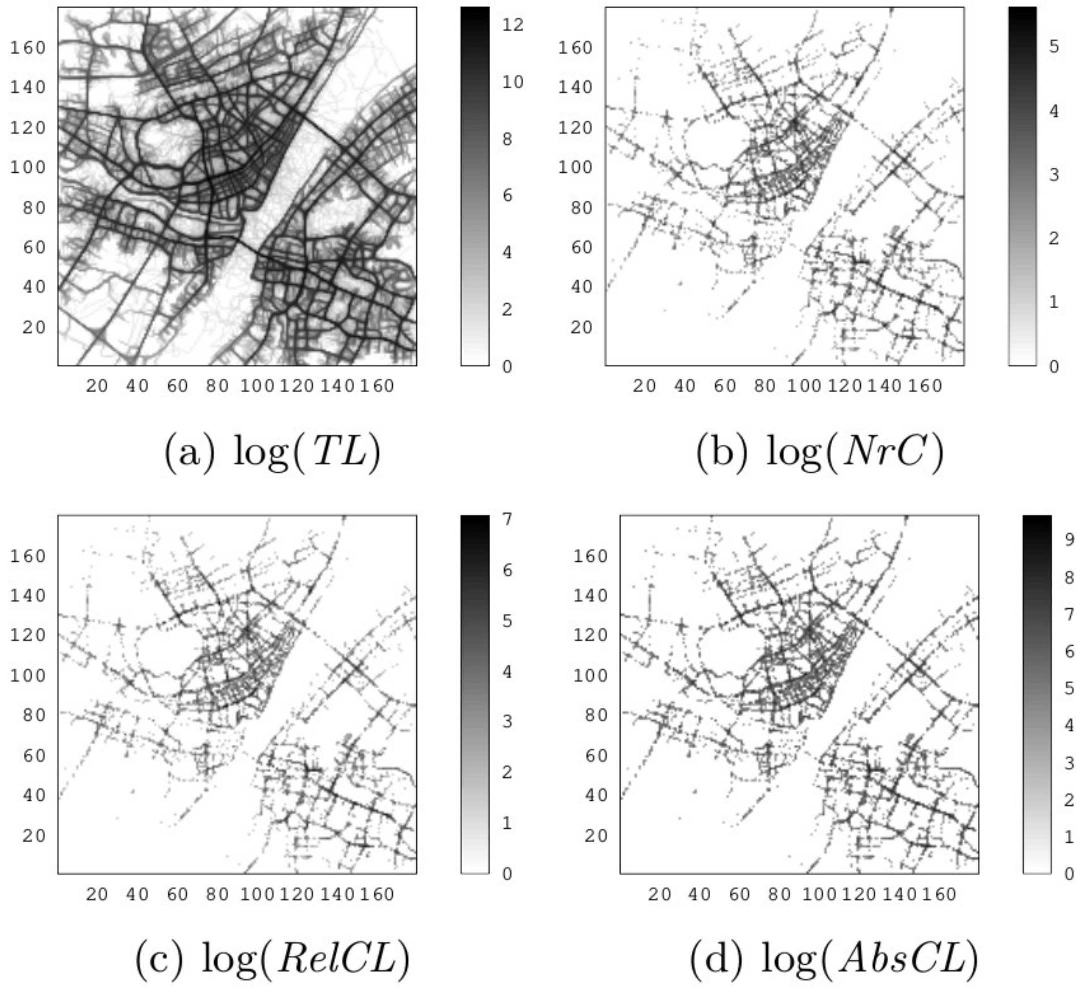


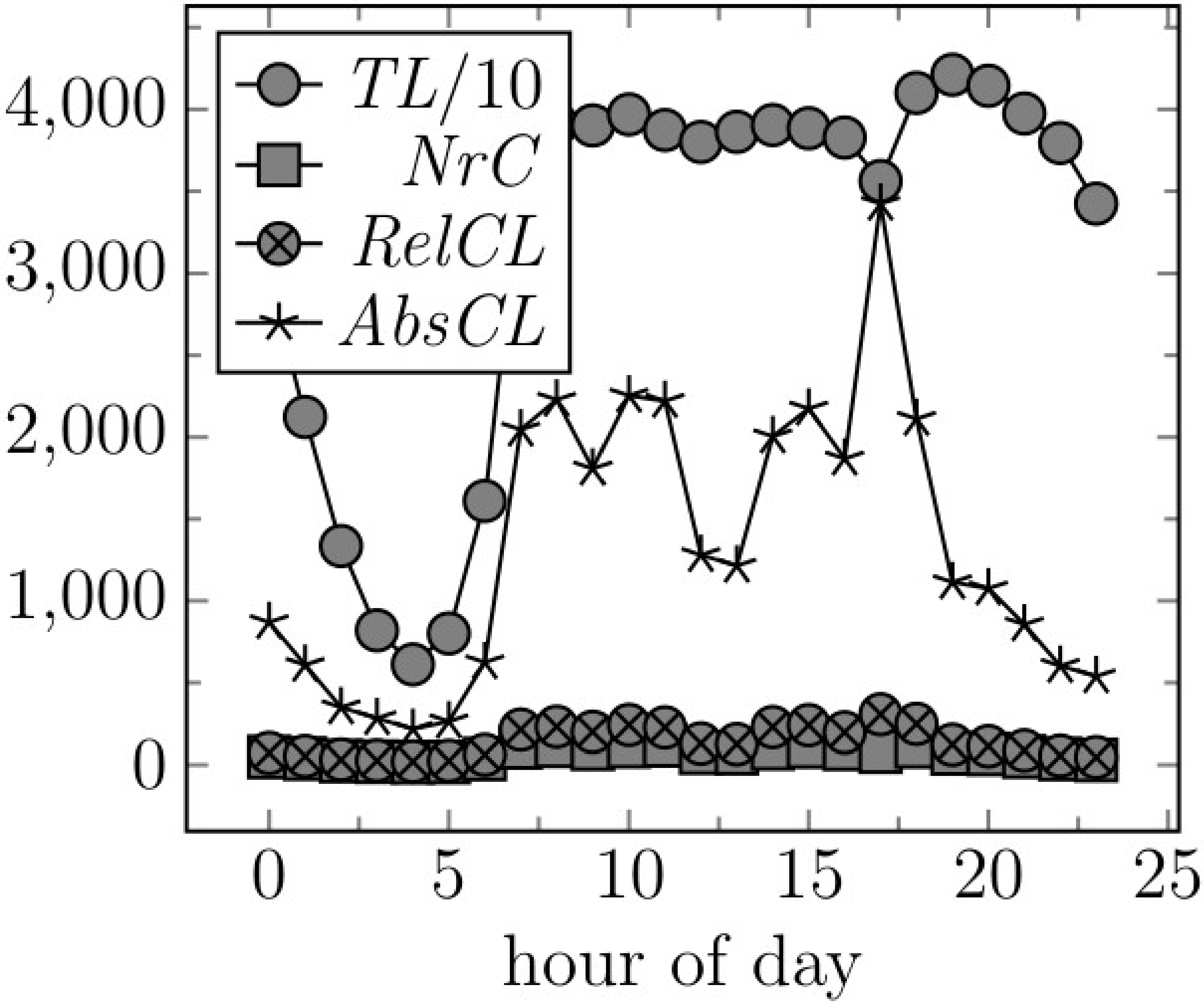
Figure 1: Spatial distribution of traffic- (TL) and congestion (NrC, $RelCL$,$AbsCL$) indicators.

Table 1: Basic distribution statistics of traffic- (TL) and congestion (NrC, $RelCL$, $AbsCL$) indicators.

Statistic	TL	NrC	$RelCL$	$AbsCL$
Minimum	2	5	5.00	17.09
Median	172.82	10	18.05	176.76
Mean	7288.3	19.69	39.48	383.22
99^{th} percentile	95934	100	367.41	3059.38
Standard deviation	19679	20.88	67.42	625.92
Maximum	291910	270	1149.2	15226

- Detections on main arteries and at intersections
- Detections are likely not the red-light periods of signaled intersections:
  - Out of the 1440 possible directional detections for a grid cell even the most frequently detected cell is only detected 270 times

Quality Assessment: Temporal Distribution



- Despite a rather constant taxi traffic levels, congestions are detected from 7am to 7pm, with morning, lunch and afternoon peaks
- At the highest level of congestion from 5-6pm the taxi traffic levels drop: perhaps both drivers and customers find this period inefficient for taxis

Quality Assessment: Congestion Clustering

- Evaluation of the strength and statistical significance of the space-time clustering of detected congestions
- Mantel test statistics: $M = \sum_{i \in E} \sum_{j \in E} X_{ij} Y_{ij}$
- Adjust for the inhomogeneity of the distribution of the underlying background population (distribution of status reports)
- 100 Monte Carlo simulations show that the detected congestions have a significantly weaker spatio-temporal clustering than random event samples from the background population.

Scalability Assessment

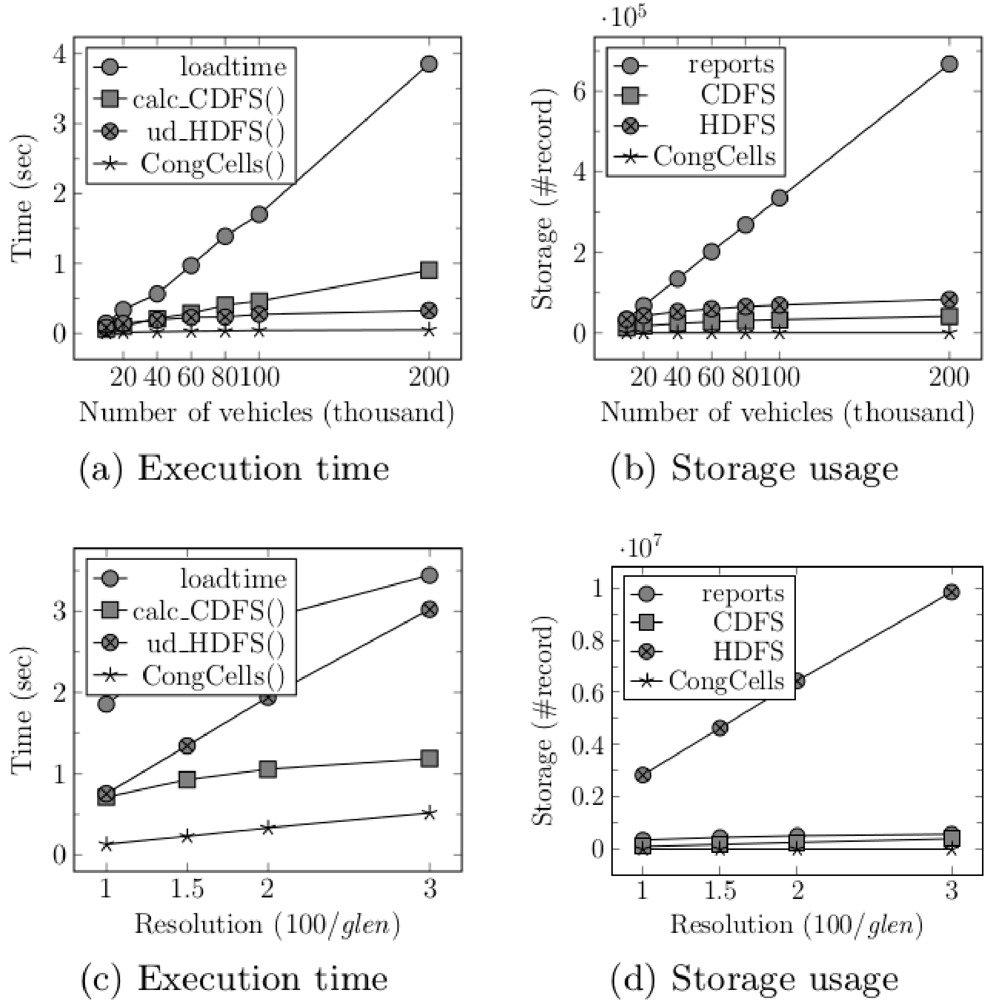


Figure 3: Execution time and space usage of different phases of the congestion detection task for varying number of vehicles and resolutions, i.e., 100/*glen*.

- Time and storage requirements of the global model scales linearly with the input size
  - Given a 60-second real-time processing limit, the system can mange approximately 60/5* 0.2K = 2.2M objects
- Time and storage requirements of the hod-projected model scales linearly (<u>not quadratically</u>) with the resolution (1/*glen*)
  - Even with millions on hod-projected HDFS, discounting load time, the system can manage 700K (@33m) – 2M (@100m) objects within the 60-second real-time limits

Outline

- Introduction
- Related work
- Method
  - Grid-based directional flow statistics
  - Directional congestion detection
  - SQL-based implementation
- Empirical evaluations
  - Quality assessment
  - Scalability assessment
- Conclusion and future work

ROYAL INSTITUTE OF TECHNOLOGY

Conclusions and Future Work

- Conclusions
  - Grid-based, time-inhomogeneous model, method, and a simple, effective, and portable SQL-implementation of the method for the detection of congestion from large FCD streams
  - Spatio-temporal distribution and clustering of the detected congestions are reasonable
  - Method and implementation scale linearly with the input size and the spatio-temporal resolution of the model
- Future work
  - Further analysis of the detected congestions
  - Use detected congestions to devise holistic congestion models
  - Road network based adaption
  - Implementation and evaluation using main-memory and stream based Big Data processing frameworks

ROYAL INSTITUTE OF TECHNOLOGY

Thank you for your attention!

Q/A?