GRID-BASED CLUSTERING OF WAZE DATA ON A RELATIONAL DATABASE

Mariana M. G. Duarte, Rebeca Schroeder , and Carmem S. Hara

Universidade Federal do Paraná, Curitiba, Brazil

Unversidade do Estado de Santa Catarina, Joinville, Brazil

INTRODUCTION

- Mobility data -> City planing
- Mobility knowledge -> Decision making

INTRODUCTION

- Big data flow -> Events continuously produced
- Increasing dataset

INTRODUCTION

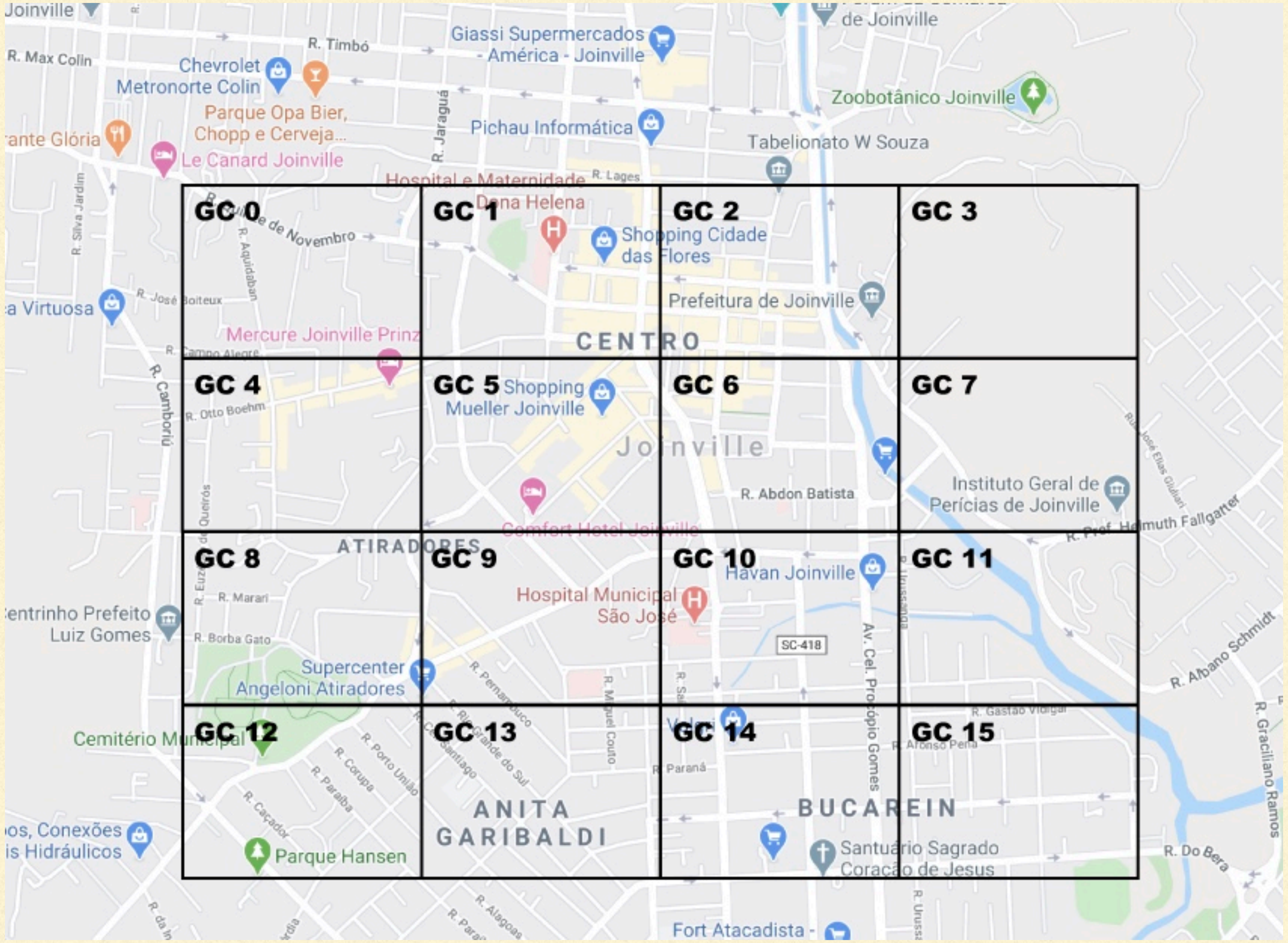
- Speed data is reported -> Stored as individual records
- Low insertion cost
- Processing spatial-temporal queries
- Avoid exhaustive search -> Index structures

INTRODUCTION

- Traditional structures adopted for spatial indexing
  - R-Trees (PostgresSQL, SQLite)
  - KD-Trees (Oracle, ExtremeDB)

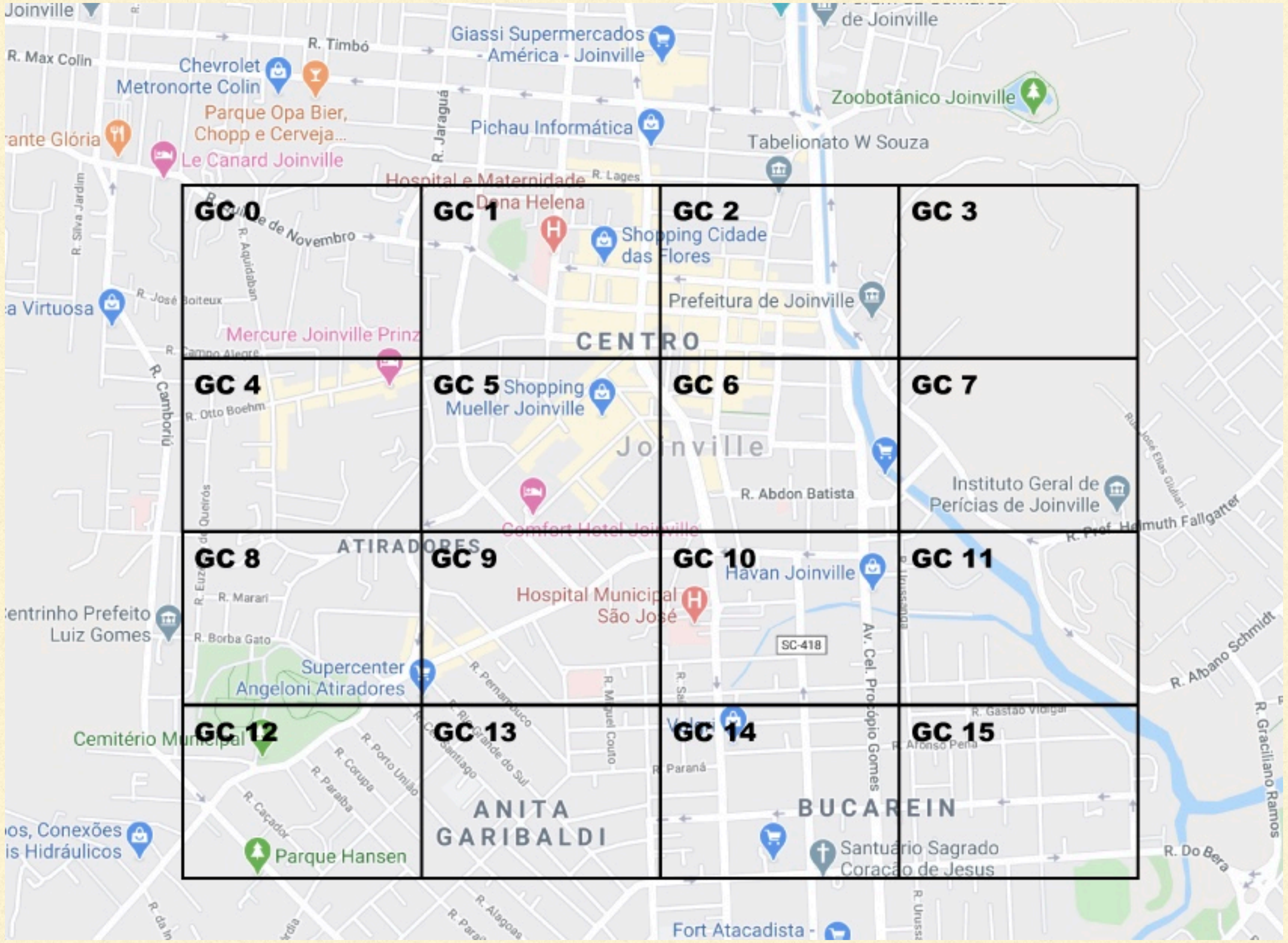
INTRODUCTION

- Partitioning of a geographic
- Creating a grid composed of juxtaposed Geographic Cells (GC)
- Eliminate the possibility of data belonging to more than one GC



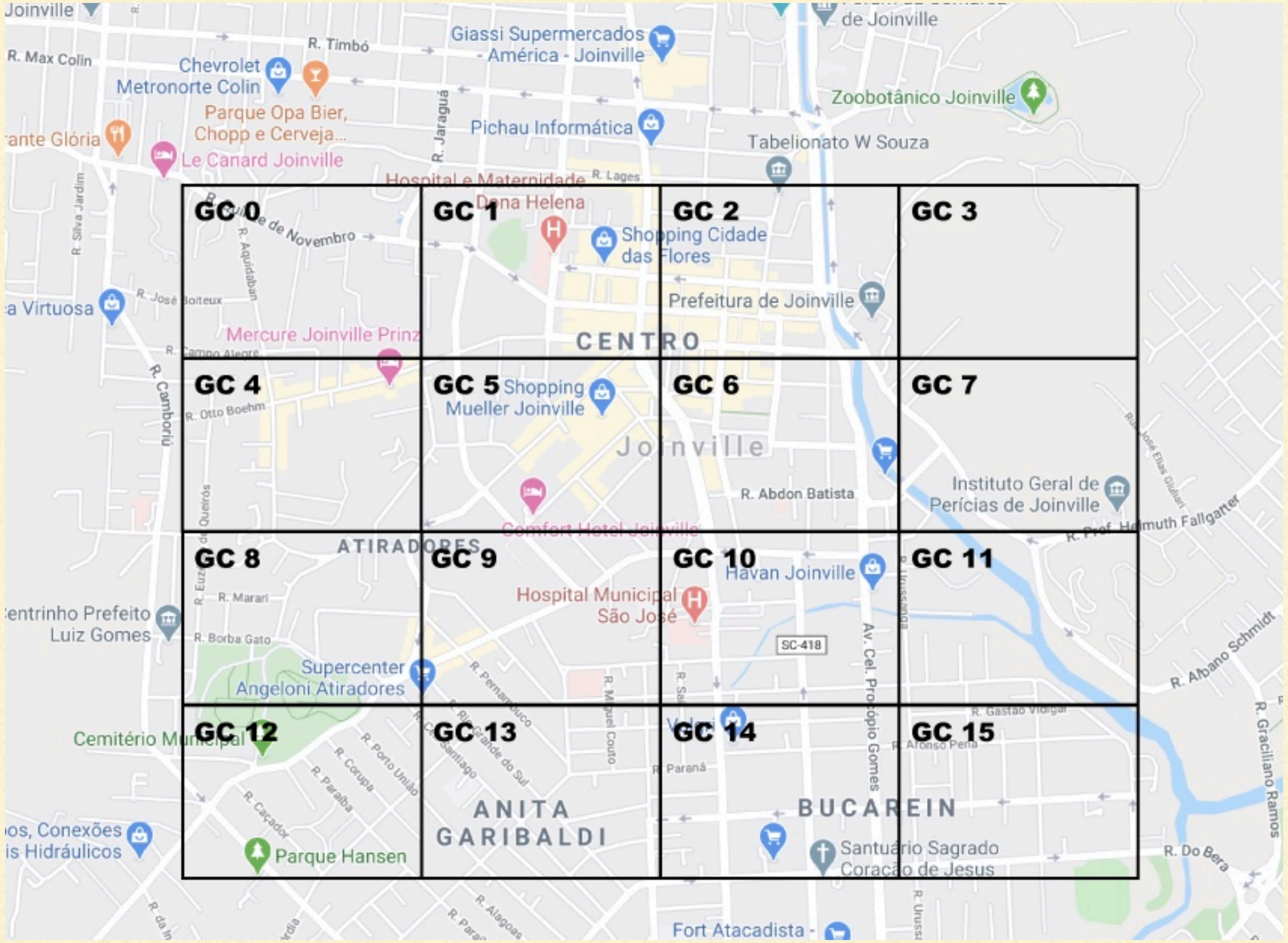
INTRODUCTION

- Not a new idea
- Basis for index structures
  - Spatial R-Tree
  - GE-Tree



INTRODUCTION

- Off-the-shelf relational database
- Traditional R-tree indexing
- Interest for developers
  - Spatial data
  - Using an existing database management system (DBMS)



INTRODUCTION

- Study case -> Data collected by Waze
  - Produced over a period of one year at Joinville, Brazil
  - Approached denoted **Waze-GC**

RELATED WORK

- Grid-based partitioning of data (Spatial R-Tree, GE-Tree, Geospark)
  - Delimited geographic area -> Matrix and predetermined number of GCs
- Spatial R-Tree -> Moving objects
  - R-tree with a grid on the leaves
  - Minimize the overlapping among minimum bounding rectangles (MBR)

RELATED WORK

- GE-Tree
  - Constant time to obtain the set of objects within GC
  - Nodes that split on demand
- GeoSpark -> Grid for a different purpose.
  - In-memory cluster -> Processing large-scale spatial data in Apache Spark
  - Grid -> Partition the data and assign cells to machines for parallel execution

RELATED WORK

- Work differs on the use of the grid
- Clustering data stored in a relational database
- Waze-GC **does not** consider
  - New indexing structures
  - Parallel processing of queries
- Waze-GC can be implemented on any relational database (RDB)

RELATED WORK

- Mobility data management
  - STIG tree
  - TQ index
- Both propose special structures for indexing mobility data
- Not based on a grid

RELATED WORK

- STIG tree -> Processed in parallel on GPUs
  - KD-tree
  - Sets of events on the leaves
- TQ index -> Predict traffic jams
  - Location index and Time index
  - Both based on hash tables

RELATED WORK

- Waze-GC works with the same type of data
- Objective: determine the impact of clustering traffic events based on a grid using native structures of a DB
- Waze-GC uses data structures that are already implemented in a RDB
    - Clustered B+ trees
    - R-trees
    - No additional structures

GRID-BASED CLUSTERING

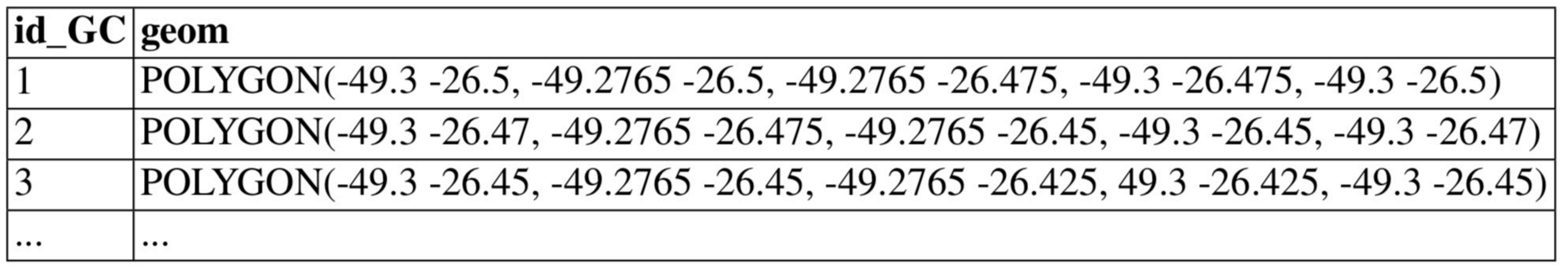
- Motivation: Clustering historic traffic events that occurred in a given area
  - Grid over the area of interest
- Possible to filter events using spatial information
- The grid-based strategy
  - Matrix of GCs of regular sizes associated
  - Set of values that represent geographical characteristics of the region

GRID-BASED CLUSTERING

- An area of interest has a MBR
  - Upper left corner coordinate (latUL, longUL)
  - Lower right corner coordinate (latLR, longLR)
  - Bounding rectangle divided into GCs
    - Non-overlapping rectangles of the same size over the area of interest

GRID-BASED CLUSTERING

- Limit of a GC -> Number rows (R) and columns (C)
- GC matrix -> GC Table
  - $R*C$ records
- This work considers linear representation of the matrix to obtain the GC's identifier (id_GC)



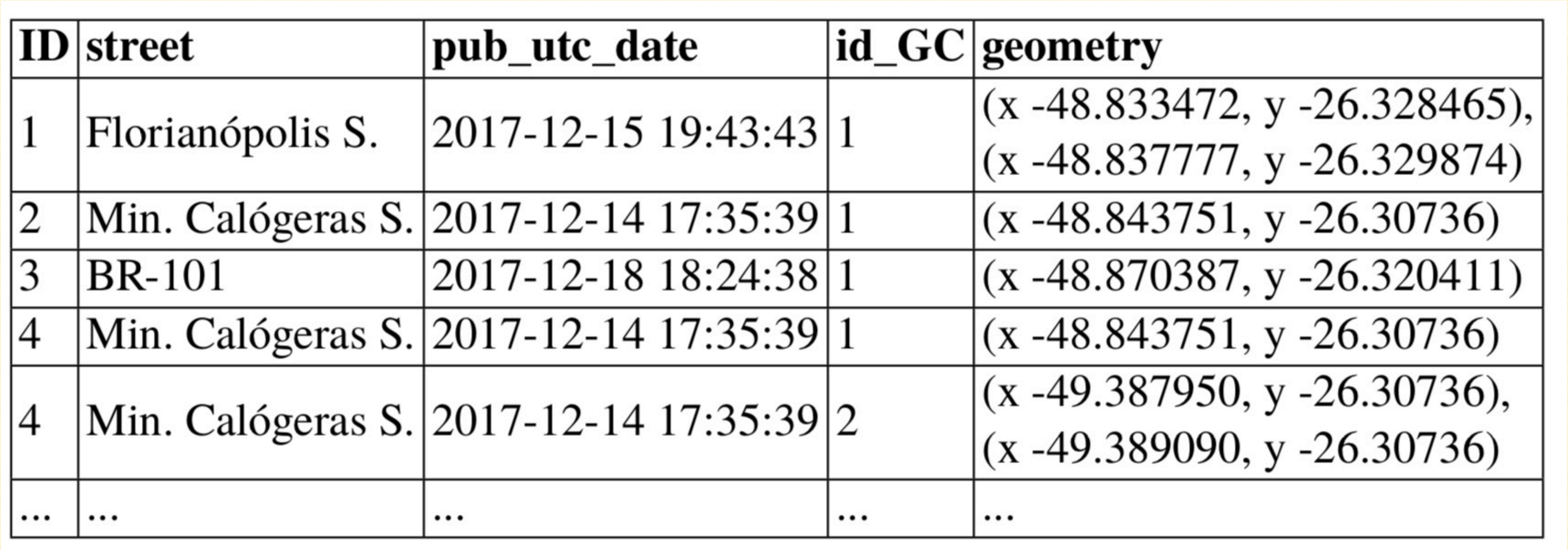
id_GC	geom
1	POLYGON(-49.3 -26.5, -49.2765 -26.5, -49.2765 -26.475, -49.3 -26.475, -49.3 -26.5)
2	POLYGON(-49.3 -26.47, -49.2765 -26.475, -49.2765 -26.45, -49.3 -26.45, -49.3 -26.47)
3	POLYGON(-49.3 -26.45, -49.2765 -26.45, -49.2765 -26.425, 49.3 -26.425, -49.3 -26.45)
...	...

GRID-BASED CLUSTERING

- Tables clustered by the attribute id_GC

- R-tree created on the geometry attribute

Waze-GC

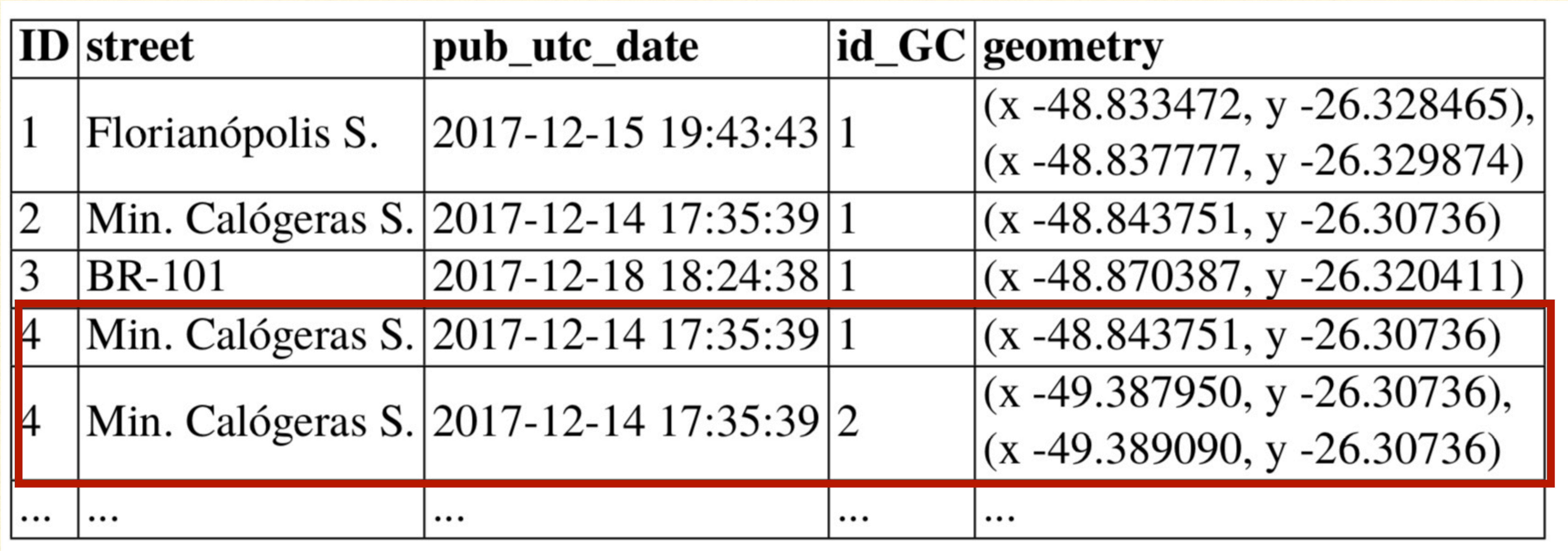


ID	street	pub_utc_date	id_GC	geometry
1	Florianópolis S.	2017-12-15 19:43:43	1	(x -48.833472, y -26.328465), (x -48.837777, y -26.329874)
2	Min. Calógeras S.	2017-12-14 17:35:39	1	(x -48.843751, y -26.30736)
3	BR-101	2017-12-18 18:24:38	1	(x -48.870387, y -26.320411)
4	Min. Calógeras S.	2017-12-14 17:35:39	1	(x -48.843751, y -26.30736)
4	Min. Calógeras S.	2017-12-14 17:35:39	2	(x -49.387950, y -26.30736), (x -49.389090, y -26.30736)
...	...	...	...	...

GRID-BASED CLUSTERING

- Primary key of the table is (event ID + id_GC)
- Jams and irregularities have a list of points
- Record r contains points p1 and p2 in different GCs it must be split

Waze-GC

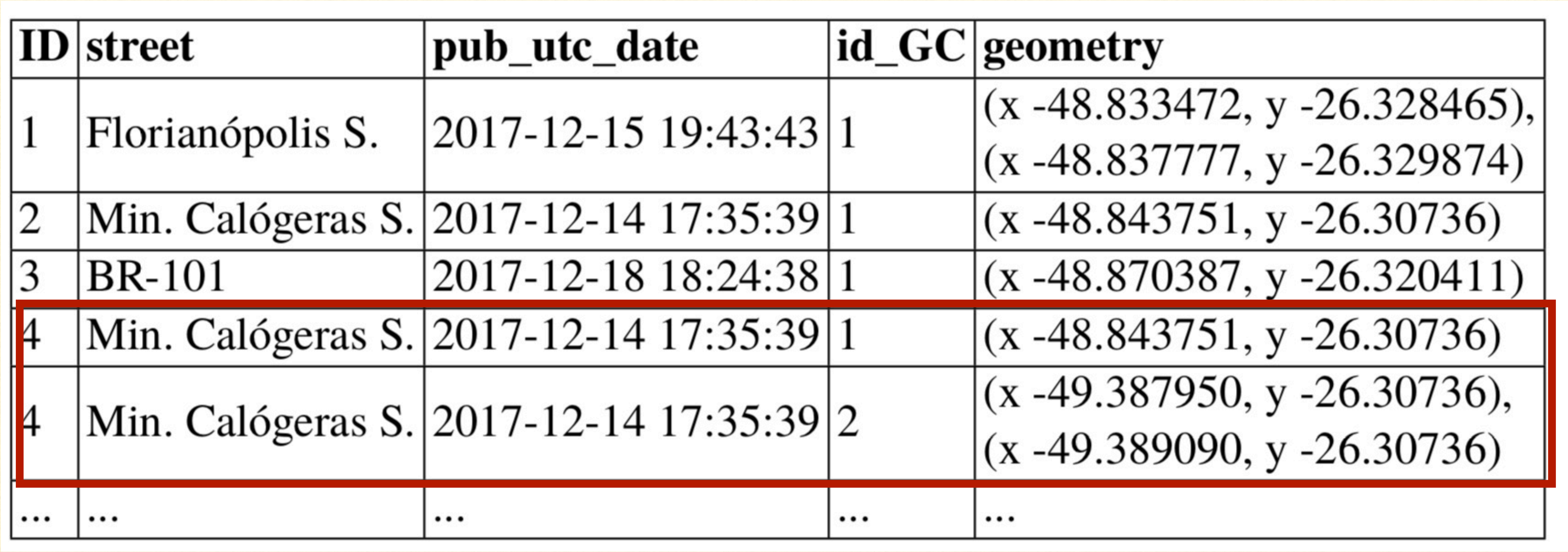


ID	street	pub_utc_date	id_GC	geometry
1	Florianópolis S.	2017-12-15 19:43:43	1	(x -48.833472, y -26.328465), (x -48.837777, y -26.329874)
2	Min. Calógeras S.	2017-12-14 17:35:39	1	(x -48.843751, y -26.30736)
3	BR-101	2017-12-18 18:24:38	1	(x -48.870387, y -26.320411)
4	Min. Calógeras S.	2017-12-14 17:35:39	1	(x -48.843751, y -26.30736)
4	Min. Calógeras S.	2017-12-14 17:35:39	2	(x -49.387950, y -26.30736), (x -49.389090, y -26.30736)
...	...	...	...	...

GRID-BASED CLUSTERING

- Two records created
- R1 contains points up to p1 and pint
- R2 contains pint and the rest
- Original event ID kept to identify points belong to same event from Waze

Waze-GC



ID	street	pub_utc_date	id_GC	geometry
1	Florianópolis S.	2017-12-15 19:43:43	1	(x -48.833472, y -26.328465), (x -48.837777, y -26.329874)
2	Min. Calógeras S.	2017-12-14 17:35:39	1	(x -48.843751, y -26.30736)
3	BR-101	2017-12-18 18:24:38	1	(x -48.870387, y -26.320411)
4	Min. Calógeras S.	2017-12-14 17:35:39	1	(x -48.843751, y -26.30736)
4	Min. Calógeras S.	2017-12-14 17:35:39	2	(x -49.387950, y -26.30736), (x -49.389090, y -26.30736)
...	...	...	...	...

GRID-BASED CLUSTERING

- Advantage of the id_GC attribute
  - Filter records related to a set of GCs
  - Join query results from different type of events from same GC
- Jams and alerts may be combined if they occurred in the same GC

EXPERIMENTAL STUDY

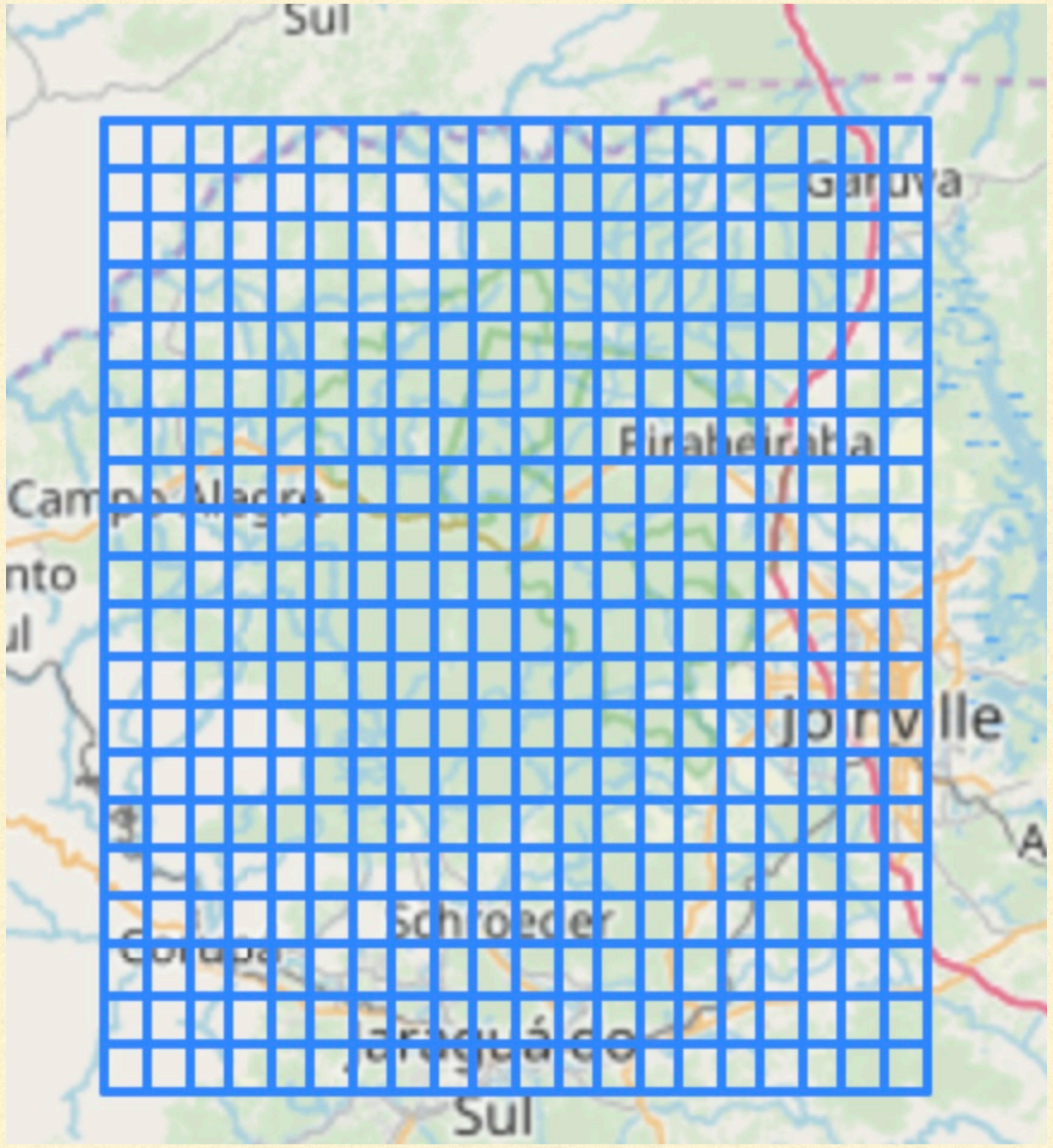
- Original Waze database
    - Postgres
    - Smart Mobility project.
    - The granting of data UDESC
    - 13 Gigabytes (GB)
    - September 2017 to September 2018

EXPERIMENTAL STUDY

- Clustered R-tree index was also defined on the geometry attribute of Waze
- Database -> Waze
- Waze-GC -> database generated by proposed approach

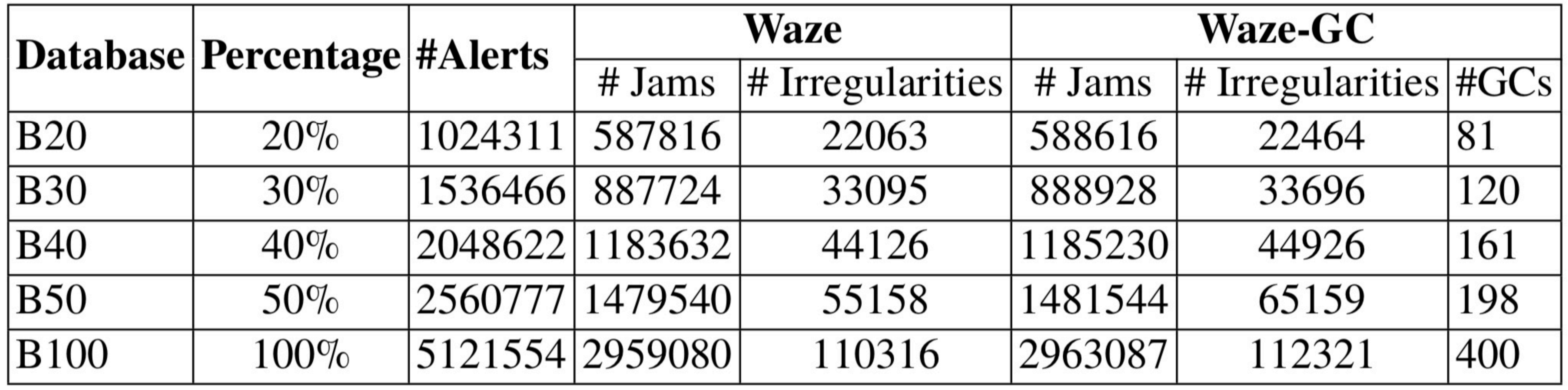
EXPERIMENTAL STUDY

- Area of interest: City of Joinville
- Grid size to 20 lines and 20 rows = 400 GCs
- Delimited by (-48.72, -26.39) (-48.92, -26.2)
- Total area has 625 km
- Cells of 1.56 km



EXPERIMENTAL STUDY

- Datasets created incrementally
- Starting with events in the central region of the city
- Larger number of records of jams and irregularities
- Set of points which may be split into different GCs



Database	Percentage	#Alerts	Waze		Waze-GC		
			# Jams	# Irregularities	# Jams	# Irregularities	#GCs
B20	20%	1024311	587816	22063	588616	22464	81
B30	30%	1536466	887724	33095	888928	33696	120
B40	40%	2048622	1183632	44126	1185230	44926	161
B50	50%	2560777	1479540	55158	1481544	65159	198
B100	100%	5121554	2959080	110316	2963087	112321	400

EXPERIMENTAL STUDY

- Waze-GC time
    - Insertions on DB
    - Stored procedure for GCs
    - Splits list of points when needed
- Process doubles the load time

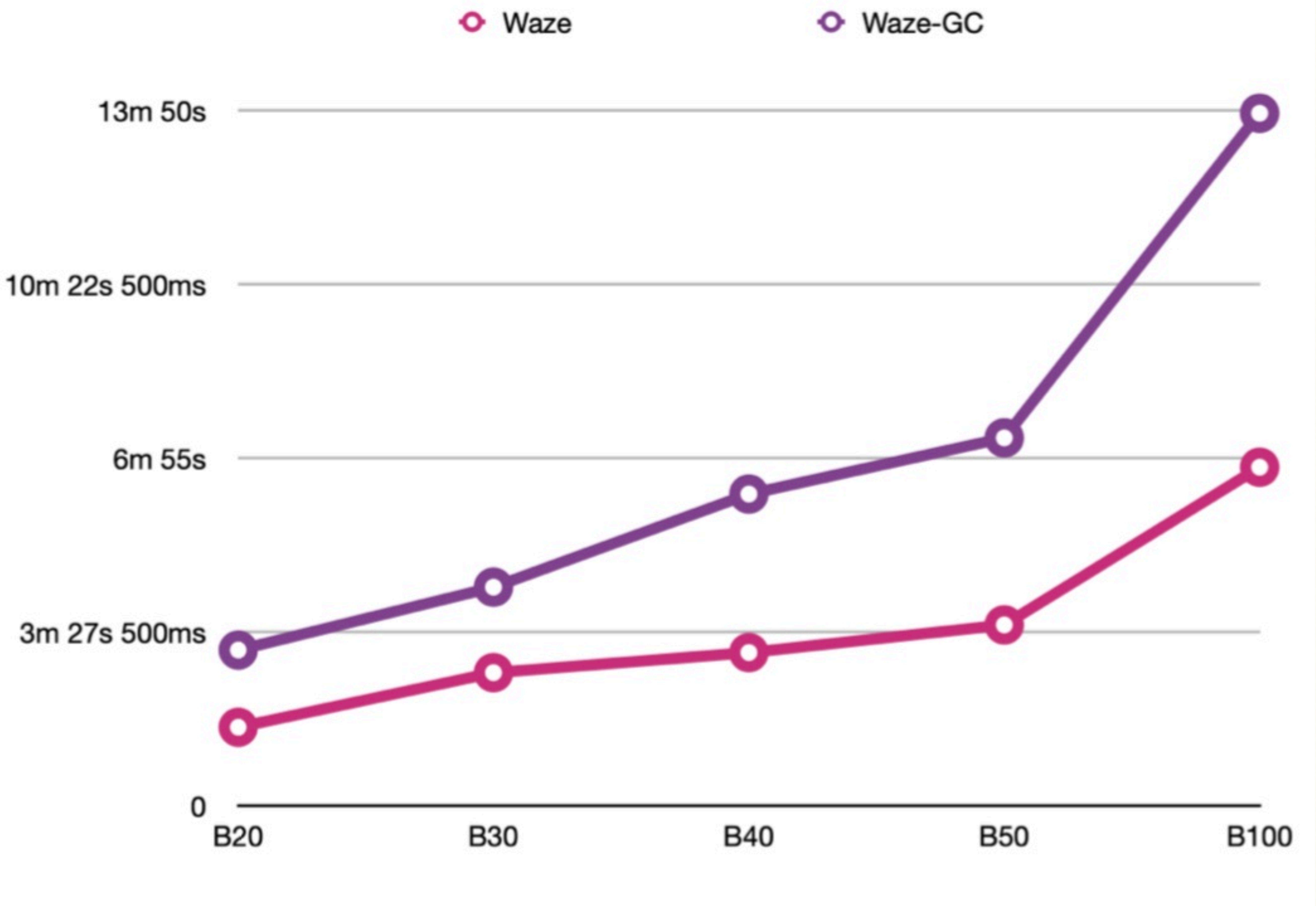


Fig. 2. Data loading time

EXPERIMENTAL STUDY

- Machine running Mac OS 10.15.2
- Dual-Core Intel Core m3 with 1.1 GHz
- 8 GB of main memory
- Two queries
  - Executed five times
  - Values consist of their average

EXPERIMENTAL STUDY

- First query
  - Analyze impact of clustering events by id_GC in Waze-GC
- "Which streets had traffic jam and alert events that occurred at exactly the same point on a street in the first seven days of 2018?"
- Waze and Waze-GC use spatial function ST_Intersects from PostGIS
- R-Tree index defined on attribute geometry -> Enhance performance of ST_Intersects

EXPERIMENTAL STUDY

- Waze-GC over preforms Waze
- Advantage by the clustering on id_GC
  - Selectivity of the join condition
  - R-tree index clustering in Waze was less effective than in Waze-GC
  - Waze-GC only compares records in the same GC

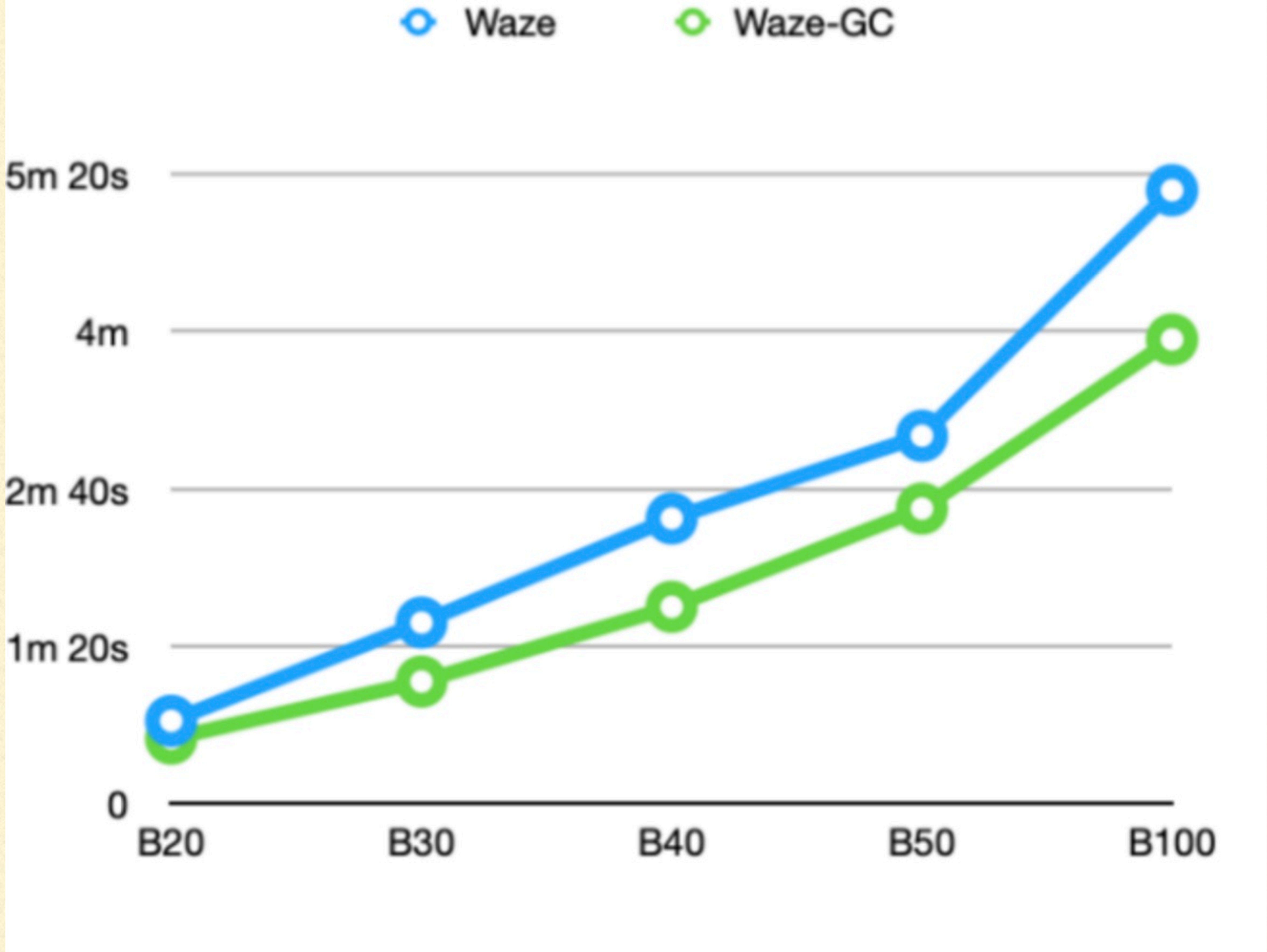
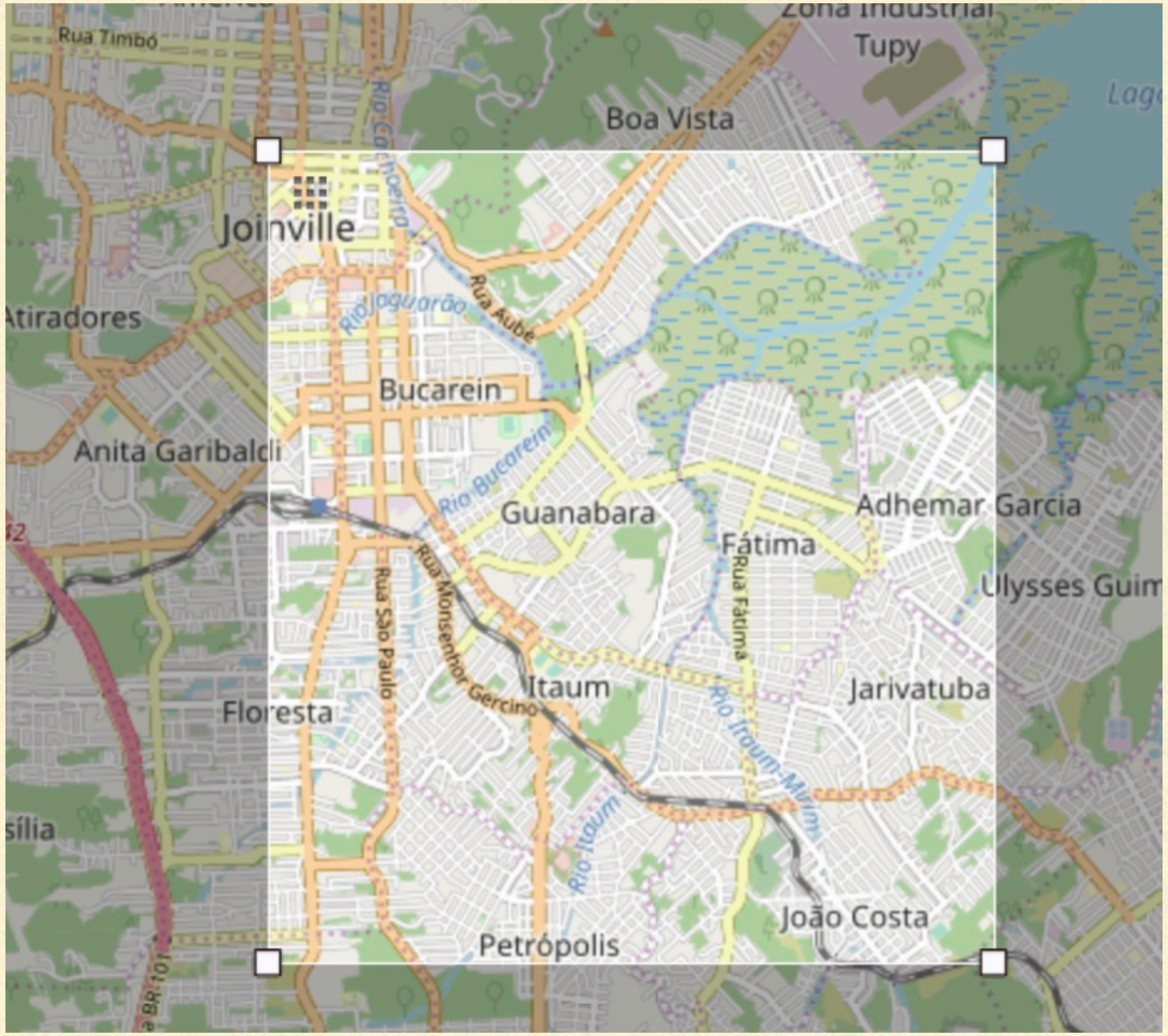


Fig. 3. Query 1 - Response Time
]]>

EXPERIMENTAL STUDY

- Second query
  - Impact of using the GC table
  - id_GC index by Waze-GC
  - "Number of traffic jams in October 2017 in an informed area of interest"
  - The size of the area of interest is 6,25 $km2$



EXPERIMENTAL STUDY

- Waze-GC statement
  - Sub-query identifies GCs in GC Table that intersect area of interest
- Returns 4 GCs for all percentages

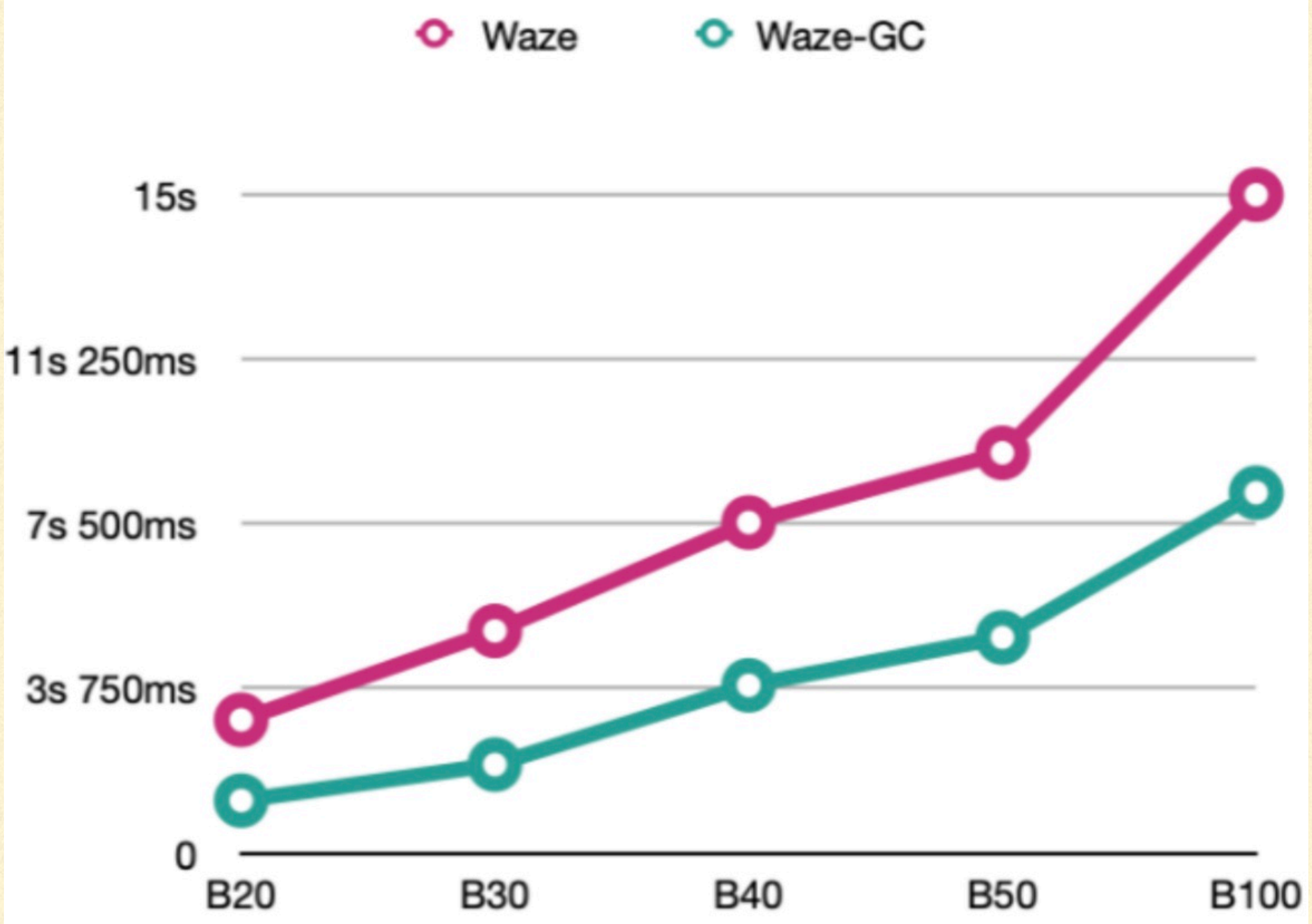


Fig. 4. Query 2 - Response Time

EXPERIMENTAL STUDY

- Advantage of Waze-GC
  - Filter id_GC -> Reduces the search space
- GC Table -> Statement more complex

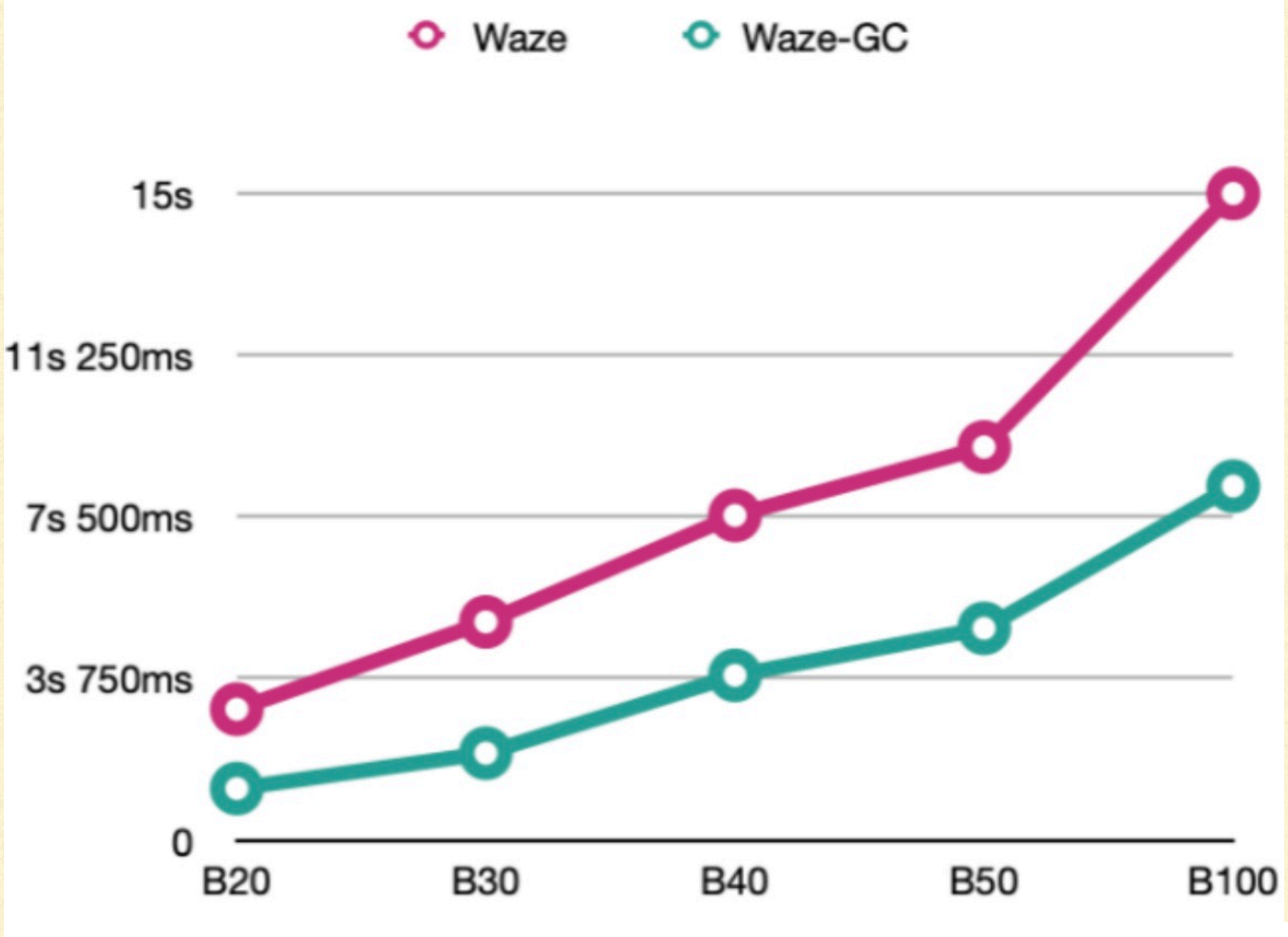


Fig. 4. Query 2 - Response Time

CONCLUSION

- Approach for partitioning an area of interest -> a grid of juxtaposed GCs
- Events that occurred in the same GC -> stored in a grouped manner
  - optimize their recovery
- Use of an off-the-shelf relational database and index structure

CONCLUSION

- Study case -> traffic events from Waze
- Two forms of data storage were tested:
  - Waze relational database -> clustered R-tree index on the geometry
  - Waze with the additional GC attribute -> clustered by GC (Waze-GC)

CONCLUSION

- Test on data sets of increasing sizes: 20%, 30%, 40%, 50% and 100%
- Advantage in query processing time of Waze-GC -> compared to Waze
- Filter on GCs reduces the search space
- More effective than an R-tree based clustering
- Query processing time is more expressive as the database size increases

CONCLUSION

- Waze-GC adds extra time on database load -> set the GC for each event
- Waze-GC has query statements more complex than usual -> GCs must be filtered
- Advantage of Waze-GC in query processing **overcomes** the shortcomings of its implementation costs

FUTURE WORK

- Determine the effect of the cell size -> better profile this approach
- Investigate its scalability
  - longer periods of time
  - larger geographic areas
- Storage alternatives to reduce the clustering overhead for inserting new records
- Files -> possible to group traffic events based on their spatial-temporal proximity
- A comprehensive comparison between alternative methods is required

REFERENCES

- 1. Doraiswamy,H.,Vo,H.T.,Silva,C.T.,Freire,J.:AGPU-BasedIndextoSupportInteractive Spatio-Temporal Queries over Historical Data. In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE). Helsinki, Finland (May 2016)
- 2. ExtremeDB:ExtremeDBDocumentation.McObject,8.0edn.(2018)
- 3. Imawan,A.,Putri,F.,Kwon,J.:TiQ:ATimelinequeryprocessingsystemoverRoadTraffic
- Data. In: 2015 IEEE International Conference on Smart City. Chengdu, China (Dec 2015)
- 4. Oracle:DataCartridgeDeveloper'sGuide.Oracle,19cedn.(2019)
- 5. PostgreSQL:PostgreSQL12.2Documentation.ThePostgreSQLGlobalDevelopmentGroup,
- 12 edn. (2020)
- 6. Rslan, E., Hameed, H.A., Ezzat, E.: Spatial R-Tree index based on grid division for query
- processing. International Journal of Database Management Systems (IJDMS) 9(6) (Dec 2017)
- 7. Shin, J., Mahmood, A., Aref, W.: An investigation of grid-enabled tree indexes for spatial query processing. In: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 169–178 (Nov 2019).
- https://doi.org/10.1145/3347146.3359384
- 8. Silva, R.: Banco De Dados Geográficos: Uma Análise Das Arquiteturas Dual (Spring) E
- Integrada (Oracle Spatial). Master's thesis, Escola Politécnica da Universidade de São Paulo,
- São Paulo, SP (2002)
- 9. SQLite:SQLiteDocumentation.SQLite,2.1.0edn.(2018)
- 10. Yu, J., Wu, J., Sarwat, M.: Geospark: A cluster computing framework for processing large- scale spatial data. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2820783.2820860

Thank you!

GRID-BASED CLUSTERING OF WAZE DATA ON A RELATIONAL DATABASE

Mariana M. G. Duarte, Rebeca Schroeder , and Carmem S. Hara
Universidade Federal do Paraná, Curitiba, Brazil
Unversidade do Estado de Santa Catarina, Joinville, Brazil