Data-Intensive Routing in Spatial Networks

Christian S. Jensen

www.cs.aau.dk/~csj

Center for Data-intensive Systems

Roadmap

- Setting: big data
- Road network travel cost modeling and computation
  - Time-varying, uncertain weights
    - Histograms
    - GMMs
- Routing
  - Stochastic skyline routing
  - Personalized routing
  - Routing based on local-driver behavior
- Closing
  - Demos, the future, challenges, acknowledgments, readings

Setting: Big Data

Hype or Substance?

- We have been pushing the boundaries for decades
  - How much data we can handle
  - How fast
  - Data integration
- Examples
  - VLDB: International Conference on Very Large Database
  - TODS: ACM Transactions on Database Systems

- So is it all hype?
  - No

Instrumentation and Digitization

- Instrumentation of reality
  - Notably, smartphones
- Digitization of processes
  - E.g., e-commerce, public services, communications, social interactions

2005 vs. 2013

Big Data

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data.**

http://www-01.ibm.com/software/data/bigdata/

Big Data Synthesis

- The result is new opportunity.
- Lots of data and unprecedented computing infrastructure combine to offer potentials for value creation from data.
- To be competitive, society and businesses must be able to create value from data.
- Data-based decisions and data-driven processes
  - Decisions based on good data beat decisions based on feelings or opinions.
- A finer granularity of services
- Entirely new services

Big Data in Routing

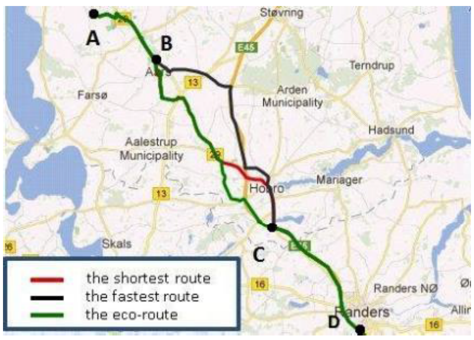
Motivation ITS

- A safer, greener, and more efficient and cost-effective transportation infrastructure
- Congestion, greater Copenhagen region
  - ~10 billion DKK/year (2004)
- Bad setting of signalized intersections in Denmark
  - ~9,3 billion DKK/year (2012)

Motivation – Eco-Routing

- The transportation sector is the second largest greenhouse gas (GHG) emitting sector and also causes substantial pollution.
  - every day, worldwide.
- The reduction of greenhouse gas (GHG) emissions from transportation is essential to combat global climate change.
  - EU: reduce GHG emissions by 30% by 2020.
  - G8: a 50% GHG reduction by 2050.
  - China: a 17% GHG reduction by 2015.
- Eco-routing can reduce vehicular impact by up to 20%.
- General context: Smart City

Motivation Eco-Weights



- The capture of the environmental costs of traversing road network edges is key to eco-routing.
  - Eco-weights are uncertain.
  - Eco-weights are time-dependent.

Time-Varying Uncertain Eco-Weights

Outline

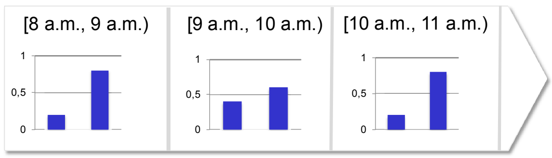
- Approach I   histograms
  - Setting
  - Framework
  - ERN building
  - GHG emissions estimation
- Approach II   GMMs
  - Setting
  - MTUG building
  - Cost estimation

Setting

- Eco Road Network G = (V, E, F)
  - V: Vertex set. Each vertex indicates a road intersection.
  - E: Edge set. Each edge indicates a road segment.
  - Function F assigns a time-dependent, uncertain eco-weight to each edge in E.
- Input
  - A set of map-matched trajectories TR.
  - An accompanying road network G' = (V, E, Null).
- Output
  - The Eco Road Network G = (V, E, F).

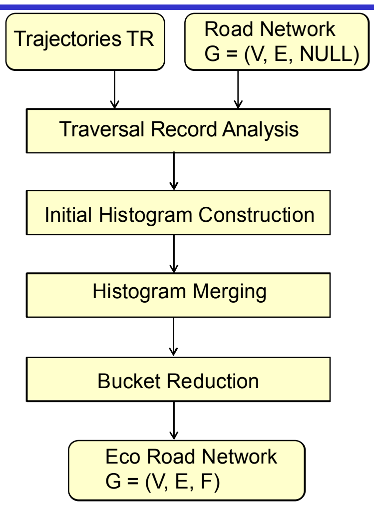
Framework

- Time-dependent uncertain histograms.
  - A vector of (*period, histogram*) tuples $<T_i, H_i>$.
  - H_i is the histogram describing the distribution of cost values observed in period T_i.



- Used to represent the eco-weights of road network edges
- Two types of compression are applied to reduce the storage space while retaining acceptable accuracy.

Framework

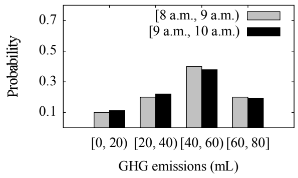


Traversal Record Analysis

- GPS records are map-matched to the corresponding edges.
- Map-matched records are transformed into traversal records.
  - A traversal record *r* = (*e, t, tt, ge*) indicates that edge *e* is traversed by a trajectory *trj* starting at time *t* and has travel time *tt* and GHG emissions *ge*.
  - The VT-micro environmental impact model is used to estimate the GHG emissions of each traversal record.

Initial Histogram Building

- Each edge is associated with a set of traversal records
- Divide the time space into intervals with equal width
  - The default value is 1 hour, (24 intervals in total).
- For each edge *e*
  - Build equi-width histograms for each time interval.
  - The number of buckets per time interval is configurable.
  - The histograms are isomorphic.

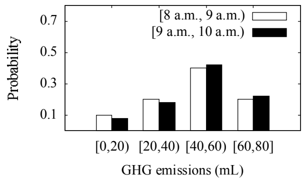


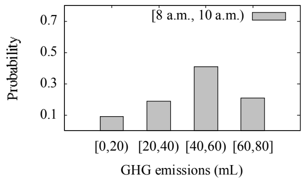
Histogram Merging

- For each edge, merge two temporally adjacent histograms if they are sufficiently similar.
- Use cosine similarity to quantify similarity.

$$sim(H_i, H_j) = \frac{V(H_i) \cdot V(H_j)}{\|V(H_i)\| \cdot \|V(H_j)\|}$$

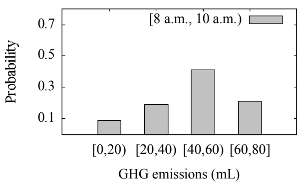
- We use a merge threshold T_{merge} to decide when to stop merging.

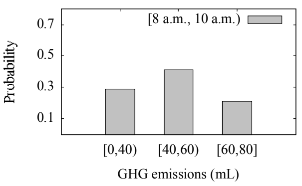




Histogram Bucket Reduction

- Further reduce the storage size of an individual histogram by merging adjacent buckets.
  - Use *SSE* to measure the merge cost (accuracy loss).
  - Merge buckets when the cost does not exceed threshold T_{red}.
  - Iteratively merge adjacent buckets in all the histograms of a road segment.

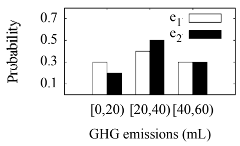


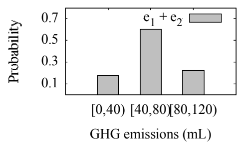


Route Cost Estimation

- For a route
  - Estimate the distribution of GHG emissions as a histogram.
  - Aggregate the histograms of the edges in the route.
- Given two histograms H_1 and H_2 for adjacent edges
  - A histogram *H' is* computed that represents the aggregated GHG emissions distribution for traversing both edges.

$$H' = H_1 + H_2$$





Outline

- Approach I – histograms
  - Setting
  - Framework
  - ERN building
  - GHG emissions estimation
- Approach II GMMs
  - Setting
  - MTUG building
  - Cost estimation

Road Network Model

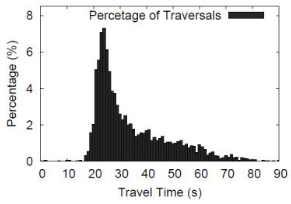
- MTUG: Multi-cost, Time-dependent, Uncertain Graph
- Assume N different costs of interest
  - Distance (DI), travel time (TT), GHG emissions (GE)
- G = (V, E, **MM**, **W**)
  - V is the vertex set, and E is the edge set.
  - **MM** = <MM$^{(1)}$ $^{(N)}$>
  - Function MM$^{(i)}$ maps an edge to the minimum and maximum i-th cost of using the edge.
    - MM$^{(TT)}$ (e_a) = (150 seconds, 500 seconds)
    - MM$^{(GE)}$ (e_b) = (10 ml, 85 ml)
  - **W** = <W$^{(1)}$ W$^{(N)}$>
  - Function W$^{(i)}$ maps an edge to a set of *(interval, random variable)* pairs of the i-th cost type.
    - W$^{(TT)}$ (e_a) = {([0:00, 7:15), N(300, 120)), ([7:15, 8:45), N(450, 100)),
    - W$^{(GE)}$ (e_b) = {([0:00, 7:00), N(30, 100)), ([7:00, 9:00), N(**50**, **80**)), …}

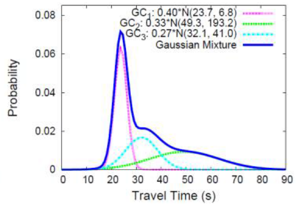
Instantiation of **MM** in an MTUG

- **MM** and **W** are instantiated using GPS records.
- GPS records are map matched to edges.
- Each edge is associated with a set of *traversal records* of the form (e, t, **C**).
  - An edge record indicates that a traversal on edge *e* at time *t* takes costs **C**, where **C** is a vector of all costs of interest**.**
  - (e_1, *8:08*, <55 seconds, 80 ml>)
  - (e_1, *9:18*, < 45 seconds , 63 ml>)
  - (e_1, 10*:10*, < 43 seconds , 60 ml>)
  - (e_1, 21*:03*, < 45 seconds , 62 ml>)
- Based on the edge records on an edge, functions MM on the edge can be instantiated.
  - $MM^{(TT)}(e_1)$ = (43 seconds, 55 seconds)
  - $MM^{(GE)}(e_1)$ = (60 ml, 80 ml)

Instantiation of **W** in an MTUG

- Partition a day into 96 15-min intervals.
- For each (edge, interval) pair, we obtain a multi-set containing the costs on the edge during the interval.
  - ms={(10 s, 3), (20 s, 10), (25 s, 20), (30 s, 10), (40 s,
- Estimate a random variable (RV) based on the multi-set.
  - Use a Gaussian Mixture Model (GMM) to represent an RV.
    - GMMs can approximate arbitrary distributions.
  - A GMM is a weighted sum of K Gaussian distributions.
    - GMM(x) =$\sum_{k=1}^{K} m_k \cdot N(x|\mu_k, \delta_k{}^2)$





Instantiation of **W** in an MTUG (cont.)

- If two RVs in two adjacent intervals are similar, we combine the two intervals into a long interval.
  - Use KL-divergence to measure the similarity between two RVs.
- Re-estimate a new RV for the long interval using the costs in the long interval.
- The whole procedure works iteratively until no RVs from consecutive intervals are similar enough to be combined.
- The long intervals along with their RVs instantiate **W.**

Route Costs in MTUG

- Given a route R_i=<r_1, r_2, …, r_X>, where r_i E is an edge.
- RC(R_i, t) indicates the costs of using route R_i at time t
  - RC(R_i, t) = <RV_{DI}, RV_{TT}, RV_{GE}> is a vector of RVs, and each RV corresponds to a travel cost.
- RV_{DI} is a deterministic value, which equals to the sum of the length of each edge in route R_i.
- RV_{TT} is the *convolution* of the corresponding travel time RV of each edge in route R_i.
  - Deciding the travel time RV of the first edge r_1 is dependent on the trip start time t.
  - Deciding the travel time RV of the k-th edge r_k is dependent on the travel time of the previous k-1 edges, which may be uncertain.
- RV_{GE} is the *convolution* of the corresponding GHG emission RV of each edge in route R_i.

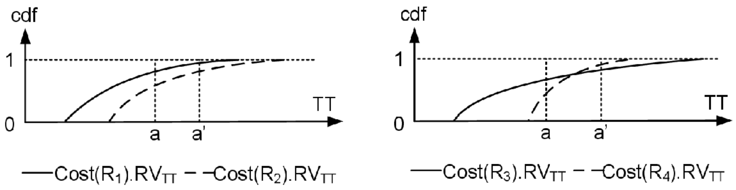
Stochastic Skyline Route Planning Under Time-Varying Uncertainty

Deterministic Skyline Routes

- Route cost: cost(R_i) = <DI, TT, GE>
  - A vector of deterministic values.
  - Each value corresponds to a travel cost.
- Dominance relationship
  - R_i dominates R_j iff all the costs of R_i are no greater than those of R_j, and there is at lest one cost of R_i is smaller than that of R_j.
- Consider multiple routes for the same source-destination.
  - R_1: 3.5 km, 230 mg, 10 min;
  - R_2: 5.1 km, 250 mg, 11 min;
  - R_3: 5.1 km, 200 mg, 12 min;
- The skyline routes are the non-dominated routes.
  - Since R_2 is dominated by R_1, R_1 and R_3 are the skyline routes.

Stochastic Dominance

- Route cost: RC(R_i) = <RV_{DI}, RV_{TT}, RV_{GE}>
  - A vector of random variables (RVs), where each RV represents the distribution of a travel cost.
- Stochastic Dominance between two RVs
  - Given two RVs X and Y, if $cdf_X(a) >= cdf_Y(a)$, for all possible value *a* in R^+, we say "X stochastically dominates Y



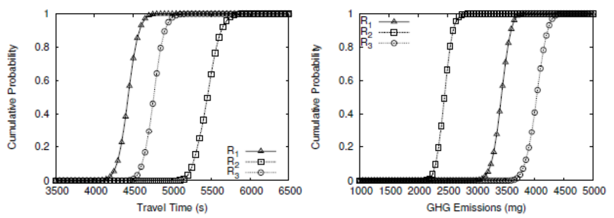
  - Cost(R_1).RV_{TT} stochastically dominates Cost(R_2). RV_{TT}.
  - No stochastic dominance between Cost(R_3). RV_{TT} and Cost(R_4). RV_{TT}.

Stochastic Skyline Routes

- Dominance between two routes R_i and R_j
  - If each RV of cost(R_i) stochastically dominates the corresponding RV of cost(R_j), then R_i dominates R_j.
- Stochastic skyline routes
  - Given a source-destination pair and a trip starting time
  - The stochastic skyline routes are the routes that are not dominated by any other routes.

Example Result

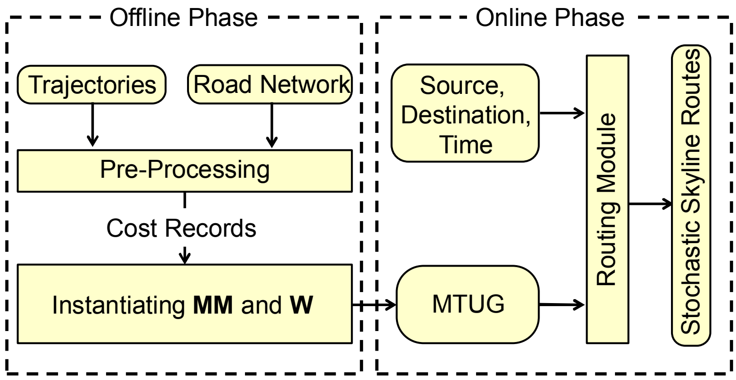
- Skyline routes R1, R2, and R3, identified by our algorithm
- R1: 94,849 m; R2: 106,216 m; R3: 91,382 m;
  - DI: R3 dominates R1 and R2.



(a) cdf of *TT* (b) cdf of *GE*

  - TT: R1 dominates R2 and R3.
  - GE: R2 dominates R1 and R3.

Framework

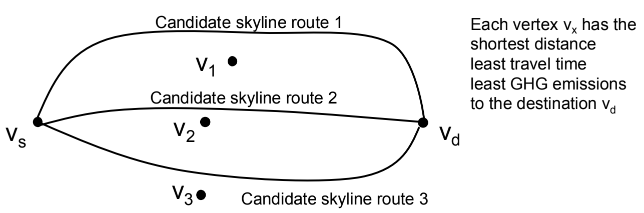


Stochastic Skyline Route Planning

- A brute force method
  - Enumerate all possible routes, compute the route costs, and check whether one route dominates another
  - Very inefficient, and works only for small road networks
- An efficient method
  - Prune some routes that cannot become skyline routes early
  - Efficient stochastic dominance checking

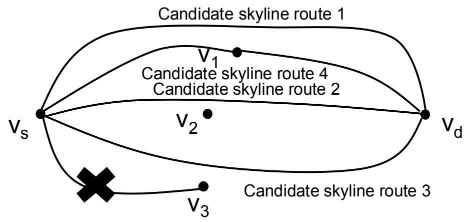
Early Pruning Strategy

- Do the following for all travel cost types of interest.
  - We use travel time as an example.
  - We maintain a graph where each edge is associated with the minimum travel time, which is recorded in **MM**.
  - From the destination, run algorithm on the graph.
    - As each vertex is associated with the minimum travel time, we get the
    - route as a candidate Skyline route.



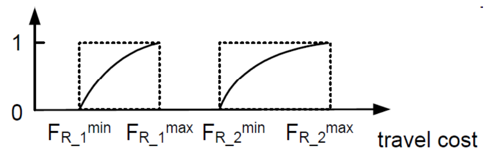
Early Pruning Strategy (cont.)

- Explore routes from source, until no more routes can be explored.
  - Estimate the least possible travel costs for a partially explored
  - If the partially explored route with its estimated least possible costs is dominated by an existing candidate skyline route, there is no need to explore the route any further.
  - Otherwise, continue exploring.
  - Update candidate skyline route if necessary.



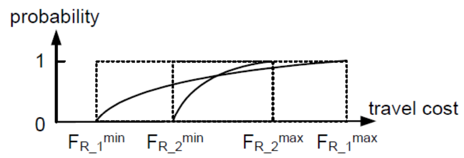
Stochastic Dominance Checking

- Naïve approach: check according to the definition of stochastic dominance.
  - For each value *a*, check whether $cdf_X(a) >= cdf_Y(a)$
- An efficient approach
  - Consider one cost type at a time
  - Compute the minimum and maximum possible travel costs of a route
- Distinguish among three cases based on the min and max travel costs of two routes
  - Disjoint case: dominance

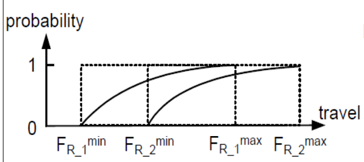


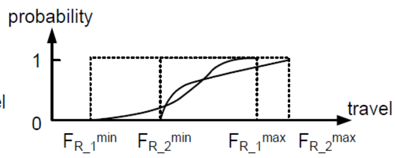
Stochastic Dominance Checking (cont.)

- Covered case: non-dominance



- Overlapping case (needs further checking)
  - Both dominance and none-dominance may occur





Summary

- Described a framework that enables stochastic skyline route planning in road networks with multiple, time-dependent, and uncertain travel costs.
- Enables eco-routing in a realistic setting.

Personalized Routing

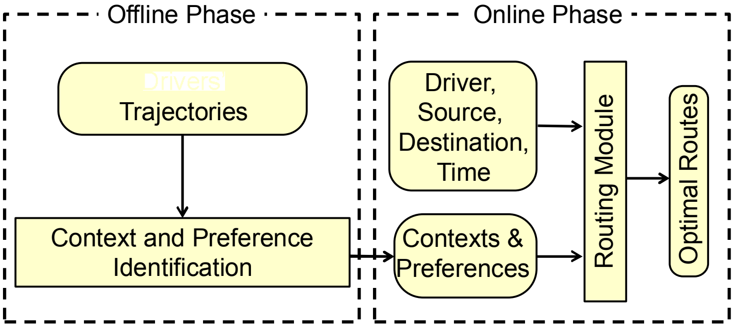
daisy

Center for Data-intensive Systems

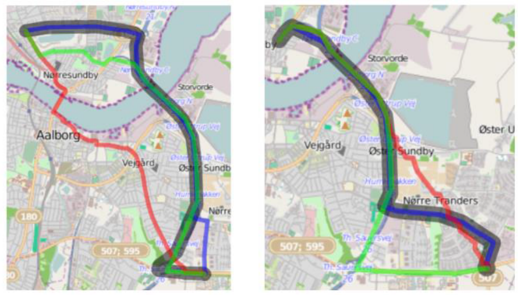
Personalized Routing

- Different drivers may take different routes because they may have quite different preferences.
- The same drivers may take different routes in different contexts.
  - Morning: try to save time to avoid being late.
  - Weekend afternoon: try to save fuel consumption.
- Challenges
  - Identify contexts for drivers and identify driving preference in each context.
  - Deal with time-dependent uncertain travel costs, e.g., travel time and fuel consumption, while considering individual drivers' driving behaviors, e.g., aggressive vs. moderate driving.

Framework



Example Results



- Dark, bold routes: actual routes used by drivers.
- Red routes: shortest routes.
- Green routes: fastest routes.
- Blue routes: predicted routes using the identified contexts and driving preferences.

skip

Vehicle Routing with User-Generated Trajectory Data

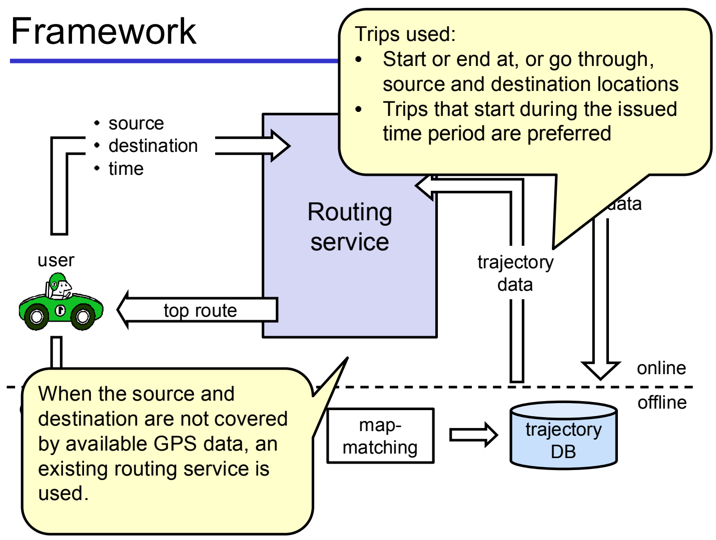
Introduction

- Local travel
  - Knowledge of the surroundings
  - Follow familiar routes
- Travel in unfamiliar surroundings to unknown destinations
  - Depend on available routing services
  - Expect that the provided route is the best
- Idea: Use GPS data to let those who travel in unfamiliar surroundings benefit from the insights of local travelers

Goal of the Study

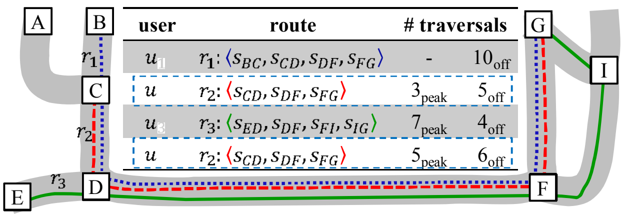
- Propose a routing framework that
  - Utilizes GPS data volunteered by local drivers
    - Exploits possibly hard-to-formalize insight into local conditions
    - Takes into account temporal variation in driver behavior
  - Recommends routes based on popularity and temporal aspects
- Evaluate the quality of proposed routes
  - Study based on trip length and pre-selected drivers
  - Quality comparison with existing routing service and route recommendation approaches

Framework



Data Preparation Methodology

- Trips that follow the same sequence of road segments are grouped into *route usage objects.*

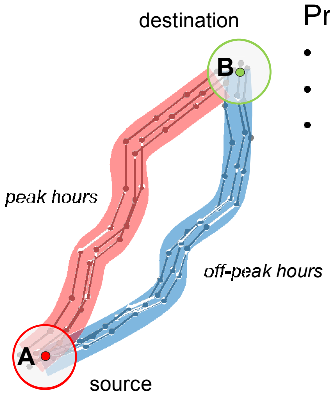


user	route	# traversals	
u_1	r_1: $\langle s_{BC}, s_{CD}, s_{DF}, s_{FG} \rangle$	-	10_{off}
u	r_2: $\langle s_{CD}, s_{DF}, s_{FG} \rangle$	3_{peak}	5_{off}
u_3	r_3: $\langle s_{ED}, s_{DF}, s_{FI}, s_{IG} \rangle$	7_{peak}	4_{off}
u	r_2: $\langle s_{CD}, s_{DF}, s_{FG} \rangle$	5_{peak}	6_{off}

- Route r_2 is taken by users u_2 and u_4 3 and 5 times during peak hours and 5 and 6 times during off-peak hours.

$$(r_2, pl, \{(peak, \{(u_2, 3), (u_4, 5)\}), (off, \{(u_2, 5), (u_4, 6)\})\})$$

Scoring of Routes



Preferred routes

- Popular among drivers
- Taken by many distinct drivers
- Popular on the time of the day and day of the week of the query

Scoring of Routes

Route preference value:

$$pref(r) = \alpha \cdot users(r) + (1 - \alpha) \cdot traversals(r)$$

distinct drivers taking the route

of traversals of the route

Final route score:

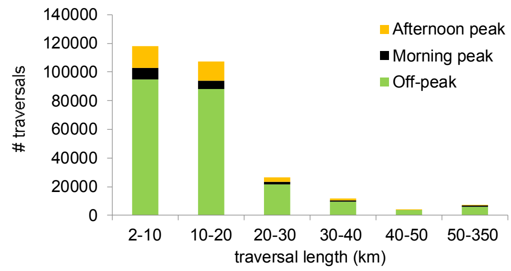
$$(r) = \beta \cdot pref^M(r) + (1 - \beta) \cdot pref^N(r)$$

Considers trips taken during the query temporal pattern

Considers trips taken during other temporal patterns

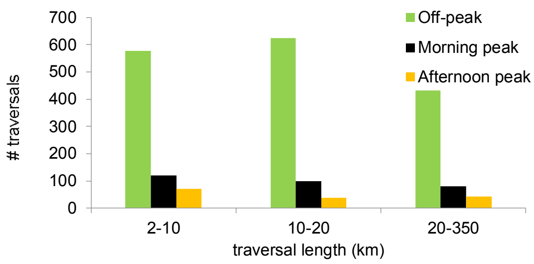
Empirical Study: Data

- Monitoring period: 2 years
- Number of drivers: 285
- Number of GPS points (raw data): ~182,700,000
- Number of trips: ~275,000
-

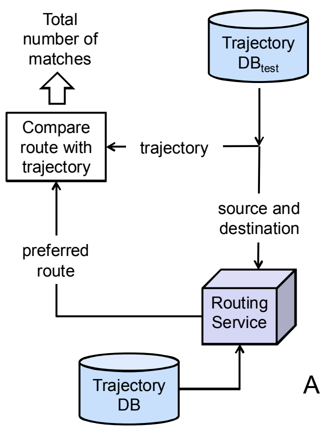


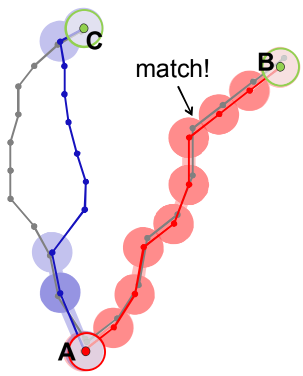
Routing Quality Evaluation: Data

For this study, we randomly selected equal amounts of trips for different trip length intervals



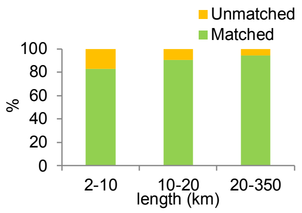
Routing Quality: Match



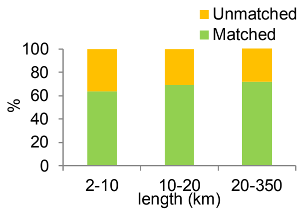


A match is identified using *LCSS*

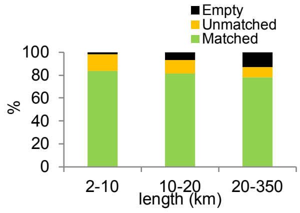
Routing Quality Evaluation: Results



Our Proposal

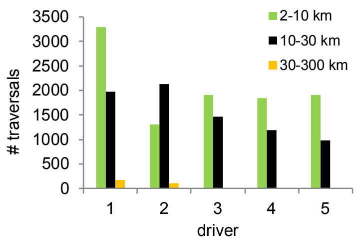


Google Directions API (top-1)

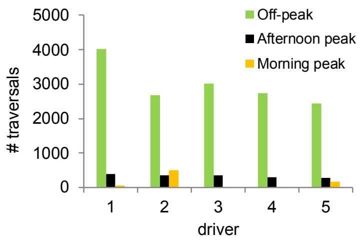


Routing Quality Evaluation: Data

For this study, we considered the five drivers with the most trips.

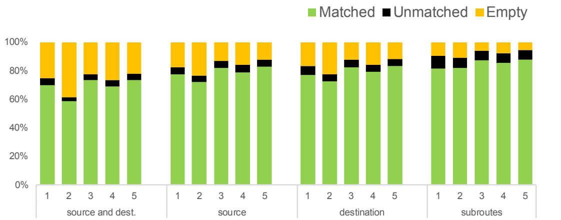


Traversals per Length



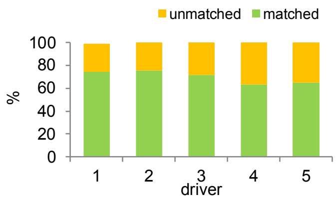
Traversals per Temporal Pattern

Routing Quality Evaluation: Results



Different Types of Route Calculation

Our proposal



Google Directions API (top-1)

Related Work

- Four existing routing techniques use [1],[2],[3],[4]
  - The road network is formed from the road segments that are covered by the trajectory data set
  - A route is formed by
    - Prioritizing parts of roads that are followed the most by a specific driver (personalized routes) [1]
    - Prioritizing parts of roads that are taken by other drivers [2],[4]
    - Possibly using sub-routes from multiple routes [3]
  - Trajectories used for scoring must contain the destination and must start and end during the provided time interval. [3]
  - Suggested routes are formed from the most popular routes or route parts in the available data set.

[1] K.-P. Chang, L.-Y. Wei, M.-Y. Yeh, and W.-C. Peng. Discovering personalized routes from trajectories. LBSN 2011, pp. 33–40
[2] Z. Chen, H. T. Shen, and X. Zhou. Discovering popular routes from trajectories. ICDE 2011, pp. 900–911
[3] W. Lou. H. Tan, L. Chen, and L.M. Ni. Finding time period-based most frequent path in big trajectory data. In SIGMOD 2013, pp. 713–724
[4] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-Drive: Driving directions based on taxi trajectories. In GIS 2010, pp. 99 108

Conclusion and Research Directions

Conclusions

- The proposed framework utilizes trajectory data collected from local drivers for routing.
- A preferred route is selected using a flexible scoring function that considers
  - The number of traversals of the route
  - The number of distinct drivers taking the route
  - The time periods when the traversals occurred
- Use of travel histories of local drivers can increase routing quality
- More details in the paper!

Research directions

- Additional aspects of the framework can be considered
  - Efficiency of route identification process (LCSS technique)
  - Inclusion of personalized routes
  - Better support for routes that are constructed from sub-routes

Closing

System of Sensors Model

- The setting may be modeled as a system of (logical) streams, one per edge.
  - Data is emitted from the stream of an edge when a vehicle traverses the edge
  - Spatial
  - Spatio-temporally correlated
  - Sparse
- Real, unlike early, envisioned smart dust applications!

Demos and Prototype Systems

- EcoTour: http://daisy.aau.dk/its/
  - Computes and compares the shortest, the fastest, and the most eco-friendly routes for arbitrary source-destination pairs in DK.
  - Best demo award at IEEE MDM 2013.
- EcoSky: http://daisy.aau.dk/its/eco/
  - Supports skyline eco-routing and personalized eco-routing
- Sheafs: http://daisy.aau.dk/its/sheaf
  - Trajectory based traffic sheafs
- Strict-Path Queries: http://daisy.aau.dk/its/spqdemo
  - Trajectory based, Strict-Path Queries
  - Trips (historical travel-time), route choice, Napoleon (road usage)
- iPark: identifying parking spaces from GPS trajectories
  - On-street parking lanes vs. parking zones

The Future

- Much more travel data
  - GPS data from vehicles
  - Inductive loop detectors, Wi-Fi/Bluetooth
  - Collective transport data, e.g., bus data,
  - Multimodal collective transport data, e.g., "Rejsekortet
- Much more connected vehicles
- New services
  - Routing
  - Safety and warnings
  - Parking, fees, insurance, road pricing
  - Car sharing, multi-modality
- Self-driving vehicles

Challenges, Examples

-
- Modeling spatio-temporal congestion from data
- Characterize the effects of events
  - Accidents, malfunctioning of traffic signals, rain, a concert
- Real-time traffic management
  - In response to current or predicted situation, actuate traffic signals and drivers (via their smartphones or navigation devices) to optimize the use of the infrastructure and driver experience
- Automated trade-off between weight level of detail and available data.
- Stochastic routing at 20 milliseconds.
- Integrate with "point" data.

Acknowledgments

- Colleagues at Aalborg University, Aarhus University, and beyond.
- The EU FP7 project, Reduction: http://www.reduction-project.eu/
- The Obel Family Foundation: http://www.obel.com/en

Readings

- V. Ceikute, C. S. Jensen: Vehicle Routing with User-Generated Trajectory Data. MDM (1) 2015
- C. Guo, B. Yang, O. Andersen, C. S. Jensen, K. Torp: EcoMark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. GeoInformatica 19(3):567-599 (2015)
- C. Guo, Bin Y., O. Andersen, C. S. Jensen, K. Torp: EcoSky: Reducing vehicular environmental impact through eco-routing. ICDE 2015:1412-1415
- B. Yang, C. Guo, Y. Ma, C. S. Jensen: Toward personalized, context-aware routing. VLDB J. 24(2):297-318 (2015)
- Y. Ma, B. Yang, C. S. Jensen: Enabling Time-Dependent Uncertain Eco-Weights For Road Networks. GeoRich@SIGMOD 2014:1:1-1:6
- B. Yang, C. Guo, C. S. Jensen, M. Kaul, S. Shang: Stochastic skyline route planning under time-varying uncertainty. ICDE 2014:136-147
- B.- Yang, M. Kaul, C. S. Jensen: Using Incomplete Information for Complete Weight Annotation of Road Networks. IEEE TKDE 26(5):1267-1279 (2014)

Readings

- C. Guo, C. S. Jensen, B. Yang: Towards Total Traffic Awareness. SIGMOD Record 43(3):18-23 (2014)
- V. Ceikute, C. S. Jensen: Routing Service Quality - Local Driver Behavior Versus Routing Services. MDM (1) 2013: 97-106
- M. Kaul, B. Yang, C. S. Jensen: Building Accurate 3D Spatial Networks to Enable Next Generation Intelligent Transportation Systems. MDM 2013: 137-146, b*est paper award.*
- Ove Andersen, Christian S. Jensen, Kristian Torp, Bin Yang: EcoTour: Reducing the Environmental Footprint of Vehicles Using Eco-routes. MDM 2013: 338-340, Demo paper, *best demo award*.
- B. Yang, C. Guo, C. S. Jensen: Travel Cost Inference from Sparse, Spatio-Temporally Correlated Time Series Using Markov Models. PVLDB 6(9): 769-780 (2013)
- C. Guo, Y. Ma, B. Yang, C. S. Jensen, Manohar Kaul: EcoMark: evaluating models of vehicular environmental impact. SIGSPATIAL/GIS 2012: 269-278