Extracting Aggregate Answer Statistics for Integration

Zainab Zolaktaf [1] Jian Xu [2] Rachel Pottinger [1]

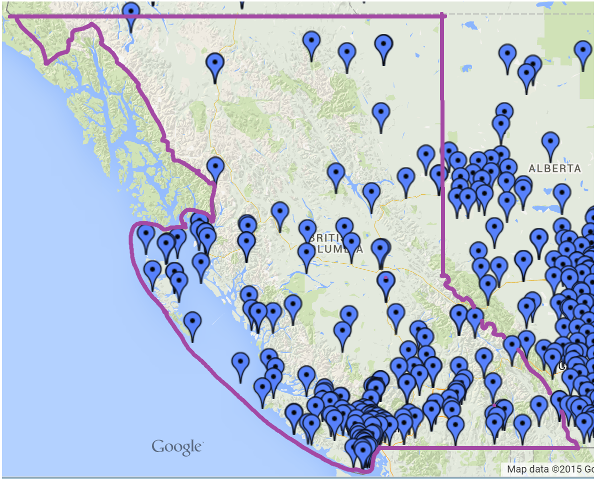
[1] Department of Computer Science
University of British Columbia

[2] Microsoft Corporation

Average high temperature across British Columbia

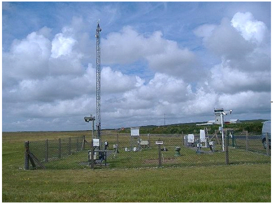
- Climate change?
  - What is the average high temperature in British Columbia for each year?
  - Averaging across the temperature over the entire province seems reasonable?
  - Query the weather stations to find the average

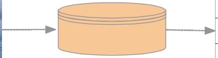
Weather stations not distributed uniformly across BC



Sources have inconsistent values for same data points

- Weather of Vancouver on 11-June-2006?

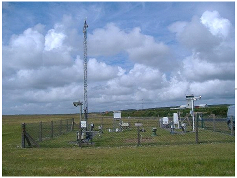




Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
Surrey	18	11-June-06
...	...	...

Sources have inconsistent values for same data points

- Weather of Vancouver on 11-June-2006?

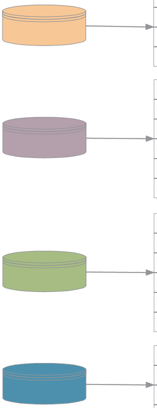


Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
Surrey	18	11-June-06
...	...	...

City	Temp	Date
Burnaby	21	06/10/06
Vancouver	22	06/11/06
Richmond	18	06/12/06
Richmond	18	06/13/06
...	...	...

Sources have different coverage and quality

- Coverage: single source contains information about a subset of objects and a subset of object attributes
- Quality: inconsistent or even conflicting values for the same object



Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
...	...	...

City	Temp	Date
Burnaby	21	06/10/06
Vancouver	22	06/11/06
Richmond	18	06/12/06
Richmond	18	06/13/06
...	...	...

City	Temp	Date	...	Total Rain
Burnaby	19	10-June-06	...	0.2
Vancouver	17	11-June-06	...	0.0
Surrey	15	11-June-06	...	0.0
Vancouver	20	12-June-06	...	1.4
...	...	...	...	...

Location	Temp	Date	Total Snow	Total Rain
Surrey	15	06/11/06	0.0	0.0
Surrey	19	06/12/06	0.0	1.2
...	...	...	...	...

Aggregate queries

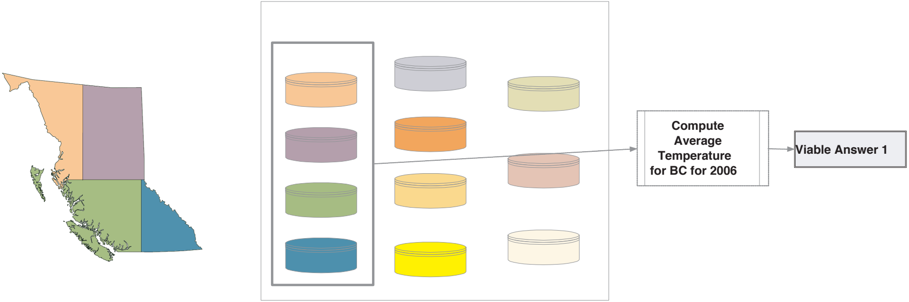
- Group set of data values and calculate informative statistics
  - Sum, Median, Avg ...
- Answering them in integration contexts
  - Requires combining sets of data that are segmented across multiple data sources
- Standard aggregation averages over all the points
  - It is incorrect!
  - Some data points have duplicates across the sources
  - The duplicates can have different values in the sources

Viable answer

- Correct aggregation requires using one value per data point
- Choosing the values from different sources will result in different answers
- Each possible answer called a viable answer
- Which set of sources and value combinations to use?

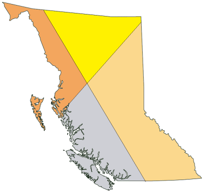
Different combinations of sources and values are possible

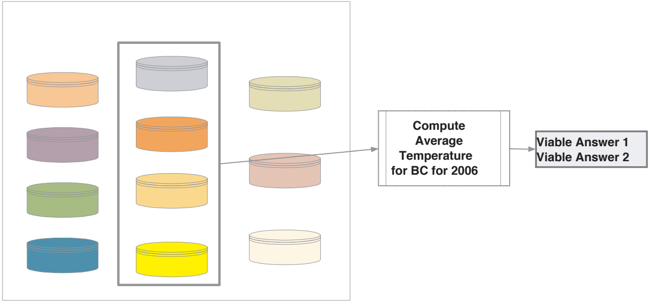
- Which set of sources and value combinations to use?



Different combinations of sources and values are possible

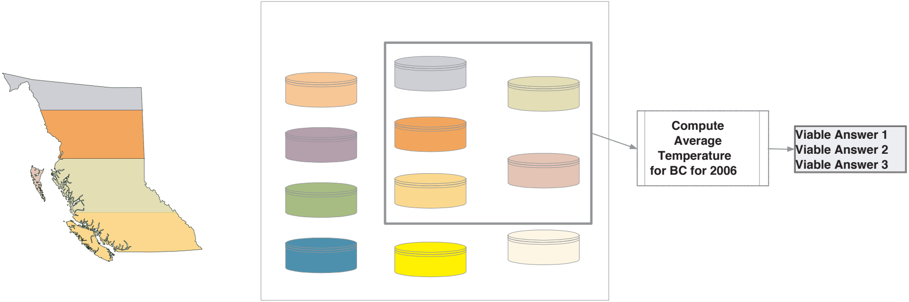
- Which set of sources and value combinations to use?





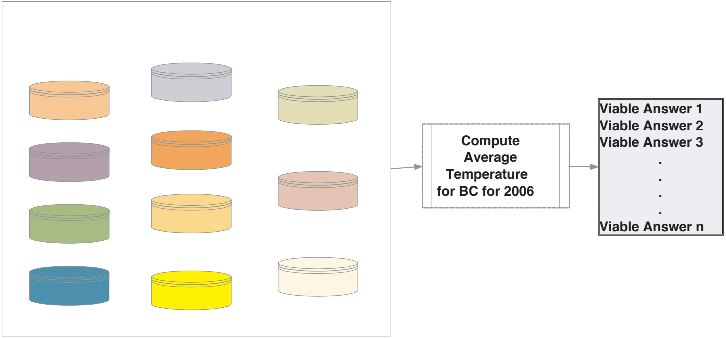
Different combinations of sources and values are possible

- Which set of sources and value combinations to use?



Different combinations of sources and values are possible

- Depending on the choice of sources and value combinations, there can be a whole range of viable answers
- Aggregate query answer is a distribution rather than a single scalar value



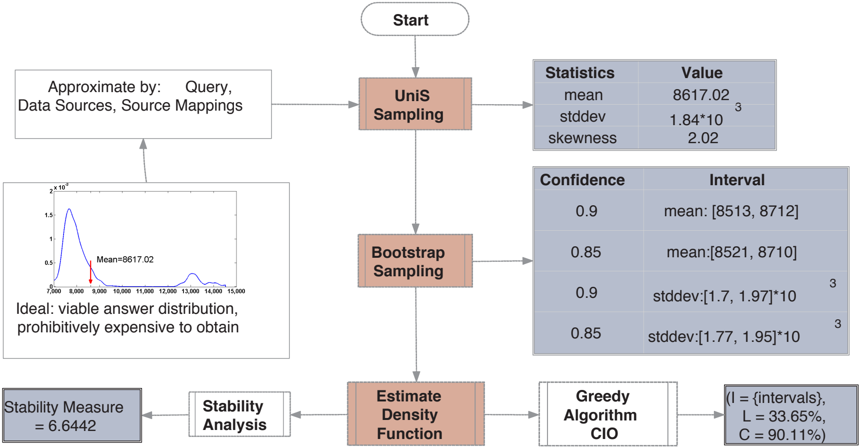
Problem formalization

- Assume meta-information regarding mapping and bindings is available
- Question: What is the viable answer distribution?
  - Enumerating all the possible value combinations is impractical
  - Estimating the exact distribution is infeasible
    - Scalability issues
    - User still has to interpret and analyze

Contributions

- We define aggregate answers as a distribution of viable answers
- We provide summary statistics for the viable answer distribution
  - Key point statistics
  - High coverage intervals
  - Stability score
- We provide algorithms for the efficient extraction of above statistics
- We verify the effectiveness of our methods using real-life and synthetic data

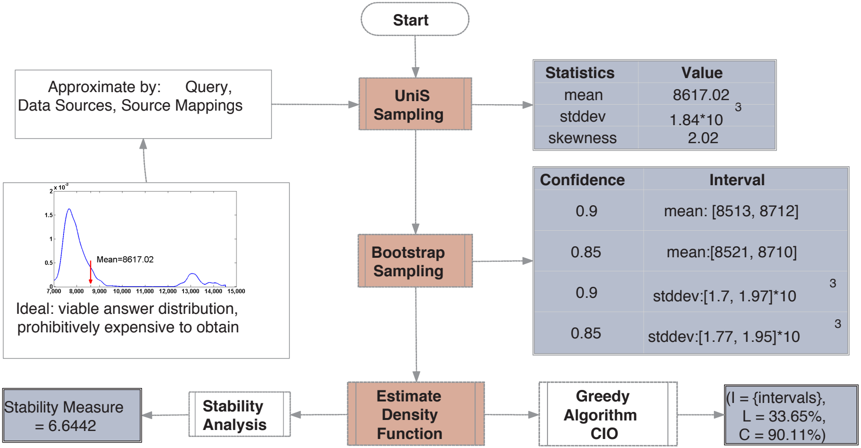
Overview


Start
UniS Sampling
Approximate by: Query, Data Sources, Source Mappings
Statistics Value
mean 8617.02
stddev 1.84*10 3
skewness 2.02
Mean=8617.02
Ideal: viable answer distribution, prohibitively expensive to obtain
Bootstrap Sampling
Confidence Interval
0.9 mean: [8513, 8712]
0.85 mean:[8521, 8710]
0.9 stddev:[1.7, 1.97]*10 3
0.85 stddev:[1.77, 1.95]*10 3
Estimate Density Function
Stability Analysis
Stability Measure = 6.6442
Greedy Algorithm CIO
(I = {intervals}, L = 33.65%, C = 90.11%)

Sampling and point statistics

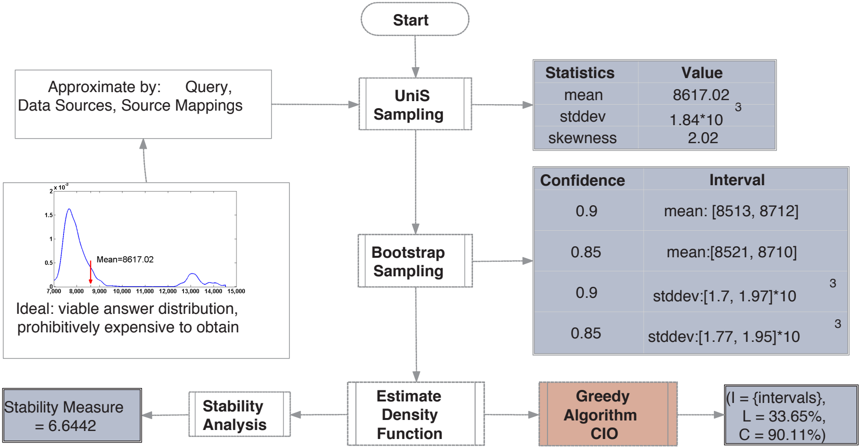
- Goal: Efficiently approximate viable answer distribution
- Sample a set of viable answers
  - No prior knowledge regarding coverage, accuracy and quality
- Sampling scheme? Uniform sampling
  - Choose sources uniformly at random
  - Stay at source until source is exhausted (all relevant components used)
- Apply bootstrap sampling and bagging
- Apply kernel density estimation

Overview

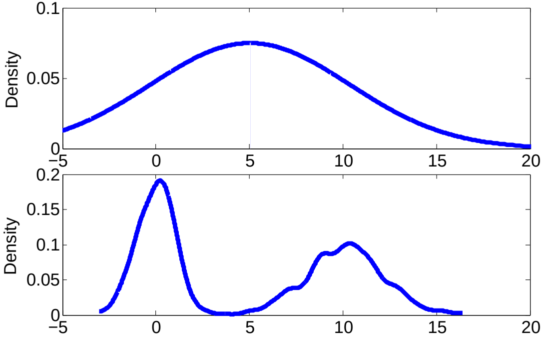


Start
Approximate by: Query, Data Sources, Source Mappings
UniS Sampling
Statistics | Value
mean | 8617.02
stddev | 1.84*10 3
skewness | 2.02
Mean=8617.02
Ideal: viable answer distribution, prohibitively expensive to obtain
Bootstrap Sampling
Confidence | Interval
0.9 | mean: [8513, 8712]
0.85 | mean:[8521, 8710]
0.9 | stddev:[1.7, 1.97]*10 3
0.85 | stddev:[1.77, 1.95]*10 3
Stability Measure = 6.6442
Stability Analysis
Estimate Density Function
Greedy Algorithm CIO
(I = {intervals}, L = 33.65%, C = 90.11%)

Overview


Start
UniS Sampling
Approximate by: Query, Data Sources, Source Mappings
Statistics Value
mean 8617.02
stddev 1.84*10 3
skewness 2.02
Mean=8617.02
Ideal: viable answer distribution, prohibitively expensive to obtain
Bootstrap Sampling
Confidence Interval
0.9 mean: [8513, 8712]
0.85 mean:[8521, 8710]
0.9 stddev:[1.7, 1.97]*10 3
0.85 stddev:[1.77, 1.95]*10 3
Estimate Density Function
Stability Analysis
Stability Measure = 6.6442
Greedy Algorithm CIO
(I = {intervals}, L = 33.65%, C = 90.11%)

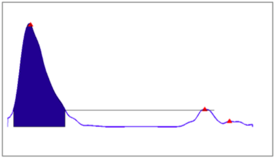
High coverage intervals and optimization



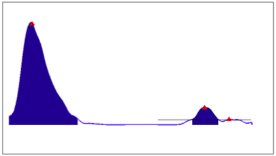
- Point statistics such as mean and variance are insufficient
- Statistics that convey shape information are needed!

High coverage intervals and optimization

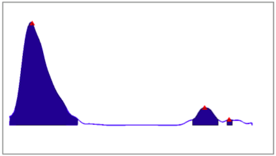
- Goal: Communicate shape information about viable answer distribution
- Greedy algorithm CIO
  - Minimize interval length so that coverage of viable answers is above a certain threshold



(a) initial high coverage interval finding

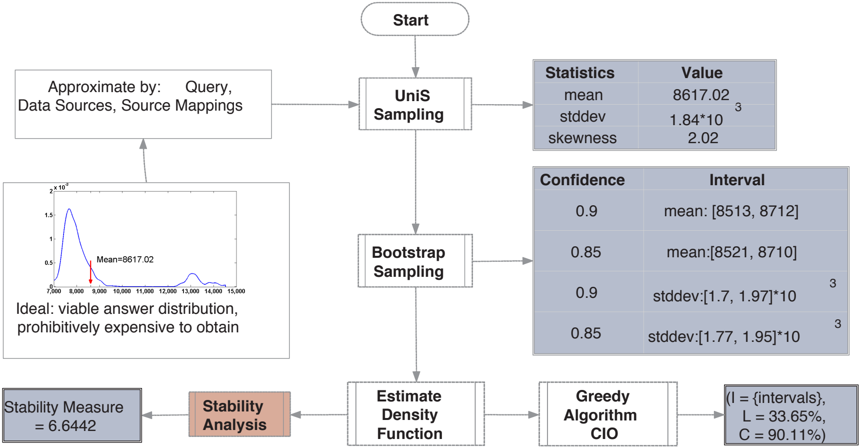


(b) intermediate



(c) high coverage intervals that are above our threshold ($\theta\%$)

Overview


Start
UniS Sampling
Approximate by: Query,
Data Sources, Source Mappings
Statistics | Value
mean | 8617.02
stddev | 1.84*10 3
skewness | 2.02
Mean=8617.02
Ideal: viable answer distribution,
prohibitively expensive to obtain
Bootstrap Sampling
Confidence | Interval
0.9 | mean: [8513, 8712]
0.85 | mean:[8521, 8710]
0.9 | stddev:[1.7, 1.97]*10 3
0.85 | stddev:[1.77, 1.95]*10 3
Estimate Density Function
Greedy Algorithm CIO
(I = {intervals},
L = 33.65%,
C = 90.11%)
Stability Analysis
Stability Measure
= 6.6442

Stability score

- Goal: How much change is caused in the estimated distribution when some sources leave?
- Quantify the change as the distance between two probability distributions
  - The original viable answer distribution
  - The viable answer distribution when some of the sources are removed
- Obtain stability analytically for the L_2 measure
- Helps prioritize re-evaluation and updating of queries need updating when sources are updated

Empirical study

- Dataset
  - Synthetic data
    - D2 - Mixtures of four Gaussians
    - D3 - Mixture of Gaussians, Cauchy and Gamma
  - Real-life data
    - C - Monthly Canadian climate data for the year 2006, from 1672 stations for 104 districts
- Aggregate query: Sum temperature data over 500 components from datasets

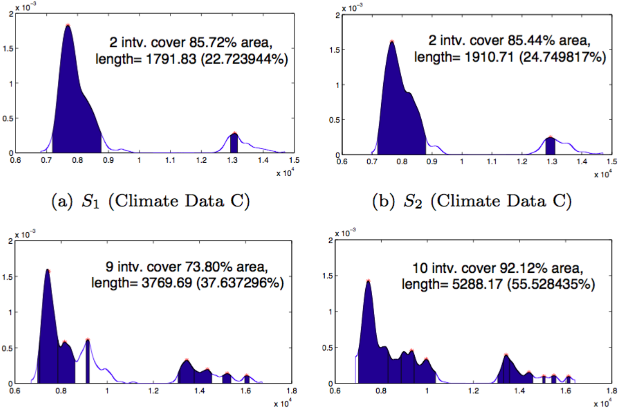
Bootstrapping vs Direct inference

- Bootstrapping helps derives tighter confidence intervals for point statistics
  - Smaller confidence intervals represent more reliable estimates
  - Improvement ratio $i_r = \frac{len(CI_{di})}{len(CI_{boot})}$

$\lvert \mathcal{S}_{uniS} \rvert$	$1-\alpha$	max i_r	avg i_r
200	0.8	4.248	2.556
200	0.9	3.309	2.119
400	0.8	2.896	2.001
400	0.9	2.293	1.655

Greedy algorithm output

- By returning dense areas, the intervals cover a small percentage of the range of data



(a) S_1 (Climate Data C) (b) S_2 (Climate Data C)

(c) S_3 (Synthetic data D3) (d) S_4 (Synthetic data D3)

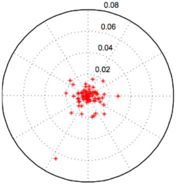
High coverage intervals

- Optimal Slicing: slice the area under the distribution into 4028 slices
- Optimal Slicing vs Greedy Algorithm CIO
  - Optimal Slicing returns tighter intervals, but does not guarantee the continuity of returned intervals

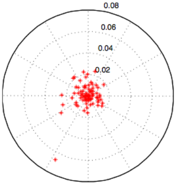
Fig	Greedy	Optimal	Cover	Greedy/Optimal
a	0.2272	0.2272	85.72%	1.0
b	0.2475	0.2475	85.44%	1.0
c	0.3764	0.2724	73.82%	1.38
d	0.5552	0.5150	92.12%	1.08

Stability score

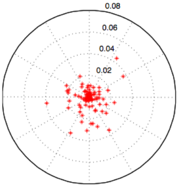
- Closer to the center, and the more dense around the center, the more stable the result



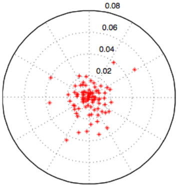
(a) S_1 L_2 score=6.5882



(b) S_2 L_2 score=6.4139



(c) S_3 L_2 score=6.4217



(d) S_4 L_2 score=6.3204

Selected related work

- Data Fusion - assumes a single true answer
  - Bleiholder, Jens, and Felix Naumann. "Data fusion." ACM Computing Surveys (CSUR) 41.1 (2008): 1.
  - Dong, Xin Luna, and Felix Naumann. "Data fusion: resolving data conflicts for integration." Proceedings of the VLDB Endowment 2.2 (2009): 1654-1655.
- Value-level heterogeneity in the Flight and Stock domain, due to sources applying different semantics
- No single true answer
  - Li, Xian, et al. "Truth finding on the deep web: is the problem solved?." Proceedings of the VLDB Endowment 6.2 (2012): 97-108.
- Applications in probabilistic databases

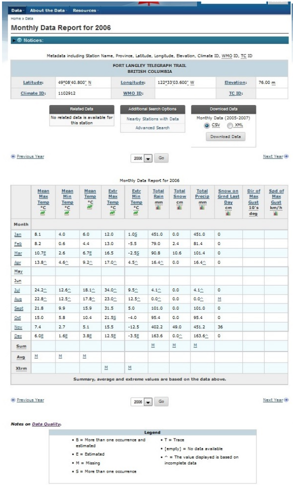
Future work

- Investigate the stability analysis
- Improve uniS sampling to consider quality and coverage
- Make inferences regarding data and sources based on non-normality of estimated viable distribution
  - Multi-modal distributions can indicate mapping problems
  - Find homogeneous sources that apply similar semantics

Contributions

- We define aggregate answers as a distribution of viable answers
- We provide summary statistics for the viable answer distribution
  - Key point statistics
  - High coverage intervals
  - Stability score
- We provide algorithms for the efficient extraction of above statistics
- We verify the effectiveness of our methods using real-life and synthetic data

Dataset



Data · About the Data · Resources

Home › Data

Monthly Data Report for 2006

Notices:

Metadata including Station Name, Province, Latitude, Longitude, Elevation, Climate ID, WMO ID, TC ID

FORT LANGLEY TELEGRAPH TRAIL BRITISH COLUMBIA					
Latitude:	49°08'40.800" N	Longitude:	122°33'03.600" W	Elevation:	76.00 m
Climate ID:	1102912	WMO ID:		TC ID:	

Related Data: No related data is available for this station

Additional Search Options: Nearby Stations with Data · Advanced Search

Download Data: Monthly Data (2005-2007) · CSV · XML · Download Data

Previous Year · 2006 · Go · Next Year

Monthly Data Report for 2006

Month	Mean Max Temp °C	Mean Min Temp °C	Mean Temp °C	Extr Max Temp °C	Extr Min Temp °C	Total Rain mm	Total Snow cm	Total Precip mm	Snow on Grnd Last Day cm	Dir of Max Gust 10's deg	Spd of Max Gust km/h
Jan	8.1	4.0	6.0	12.0	1.0S	451.0	0.0	451.0	0		
Feb	8.2	0.6	4.4	13.0	-5.5	79.0	2.4	81.4	0		
Mar	10.7E	2.6	6.7E	16.5	-2.5S	90.8	10.6	101.4	0		
Apr	13.8^	4.6^	9.2^	17.0^	4.5^	16.4^	0.0	16.4^	0		
May											
Jun											
Jul	24.2^	12.6^	18.1^	34.0^	9.5^	4.1^	0.0	4.1^	0		
Aug	22.8^	12.5^	17.8^	23.0^	12.5^	0.0^	0.0	0.0^	M		
Sept	21.8	9.9	15.9	31.5	5.0	101.0	0.0	101.0	0		
Oct	15.0	5.8	10.4	21.5S	-4.0	95.4	0.0	95.4	0		
Nov	7.4	2.7	5.1	15.5	-12.5	402.2	49.0	451.2	36		
Dec	6.0E	1.6E	3.8E	12.5E	-3.5E	163.6	0.0^	163.6^	0		
Sum						M	M	M			
Avg	M	M	M								
Xtrm				M	M						

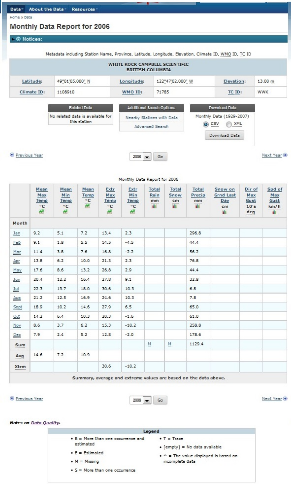
Summary, average and extreme values are based on the data above.

Previous Year · 2006 · Go · Next Year

Notes on Data Quality.

Legend

- B = More than one occurrence and estimated
- E = Estimated
- M = Missing
- S = More than one occurrence
- T = Trace
- [empty] = No data available
- ^ = The value displayed is based on incomplete data



Data · About the Data · Resources

Home › Data

Monthly Data Report for 2006

Notices:

Metadata including Station Name, Province, Latitude, Longitude, Elevation, Climate ID, WMO ID, TC ID

WHITE ROCK CAMPBELL SCIENTIFIC BRITISH COLUMBIA					
Latitude:	49°01'05.000" N	Longitude:	122°47'02.000" W	Elevation:	13.00 m
Climate ID:	1108910	WMO ID:	71785	TC ID:	WWK

Related Data: No related data is available for this station

Additional Search Options: Nearby Stations with Data · Advanced Search

Download Data: Monthly Data (1929-2007) · CSV · XML · Download Data

Previous Year · 2006 · Go · Next Year

Monthly Data Report for 2006

Month	Mean Max Temp °C	Mean Min Temp °C	Mean Temp °C	Extr Max Temp °C	Extr Min Temp °C	Total Rain mm	Total Snow cm	Total Precip mm	Snow on Grnd Last Day cm	Dir of Max Gust 10's deg	Spd of Max Gust km/h
Jan	9.2	5.1	7.2	13.4	2.3			296.8			
Feb	9.1	1.8	5.5	14.5	-4.5			44.4			
Mar	11.4	3.8	7.6	16.8	-2.2			56.2			
Apr	13.8	6.2	10.0	21.3	2.3			76.8			
May	17.6	8.6	13.2	26.8	2.9			44.4			
Jun	20.4	12.2	16.4	27.8	9.1			32.8			
Jul	22.3	13.7	18.0	30.6	10.3			6.8			
Aug	21.2	12.5	16.9	24.6	10.3			7.8			
Sept	18.9	10.2	14.6	27.9	6.5			65.0			
Oct	14.2	6.4	10.3	20.3	-1.6			61.0			
Nov	8.6	3.7	6.2	15.3	-10.2			258.8			
Dec	7.9	2.4	5.2	12.8	-2.0			178.6			
Sum						M	M	1129.4			
Avg	14.6	7.2	10.9								
Xtrm				30.6	-10.2						

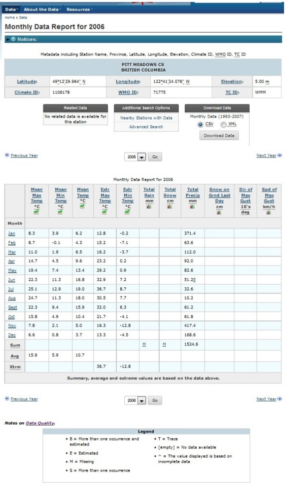
Summary, average and extreme values are based on the data above.

Previous Year · 2006 · Go · Next Year

Notes on Data Quality.

Legend

- B = More than one occurrence and estimated
- E = Estimated
- M = Missing
- S = More than one occurrence
- T = Trace
- [empty] = No data available
- ^ = The value displayed is based on incomplete data



Data · About the Data · Resources

Home › Data

Monthly Data Report for 2006

Notices:

Metadata including Station Name, Province, Latitude, Longitude, Elevation, Climate ID, WMO ID, TC ID

PITT MEADOWS CS BRITISH COLUMBIA					
Latitude:	49°12'29.964" N	Longitude:	122°41'24.076" W	Elevation:	5.00 m
Climate ID:	1106178	WMO ID:	71775	TC ID:	WMM

Related Data: No related data is available for this station

Additional Search Options: Nearby Stations with Data · Advanced Search

Download Data: Monthly Data (1993-2007) · CSV · XML · Download Data

Previous Year · 2006 · Go · Next Year

Monthly Data Report for 2006

Month	Mean Max Temp °C	Mean Min Temp °C	Mean Temp °C	Extr Max Temp °C	Extr Min Temp °C	Total Rain mm	Total Snow cm	Total Precip mm	Snow on Grnd Last Day cm	Dir of Max Gust 10's deg	Spd of Max Gust km/h
Jan	8.3	3.9	6.2	12.8	-0.2			371.4			
Feb	8.7	-0.1	4.3	15.2	-7.1			63.6			
Mar	11.0	1.9	6.5	16.2	-3.7			112.0			
Apr	14.7	4.5	9.6	23.2	0.2			92.0			
May	19.4	7.4	13.4	29.2	0.9			82.6			
Jun	22.3	11.3	16.8	32.9	7.2			51.2E			
Jul	25.1	12.9	19.0	36.7	8.7			32.6			
Aug	24.7	11.3	18.0	30.5	7.7			10.2			
Sept	22.3	9.4	15.9	32.0	6.3			61.2			
Oct	15.8	4.9	10.4	21.7	-4.1			61.8			
Nov	7.8	2.1	5.0	16.3	-12.8			417.4			
Dec	6.6	0.8	3.7	13.3	-4.5			168.6			
Sum						M	M	1524.6			
Avg	15.6	5.9	10.7								
Xtrm				36.7	-12.8						

Summary, average and extreme values are based on the data above.

Previous Year · 2006 · Go · Next Year

Notes on Data Quality.

Legend

- B = More than one occurrence and estimated
- E = Estimated
- M = Missing
- S = More than one occurrence
- T = Trace
- [empty] = No data available
- ^ = The value displayed is based on incomplete data

Preliminaries

- Bootstrap sampling and bagging
- Kernel density estimation
- Distance measures for distributions

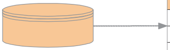
Finding high coverage intervals - optimization approach

Given a density function $f_X^{\mathcal{D}}$ for a distribution defined on a finite range, a coverage threshold $0 \leq \theta \leq 1$, and a constant t representing the number of modes, the CIO problem finds k intervals $I_1, I_2, \ldots, I_k$ where $k \leq t$, to minimize $\sum_{i=1}^{k} |I_i|$ subject to $\sum_{i=1}^{k} \int_{I_i} f_X^{\mathcal{D}}(x)dx \geq \theta$.

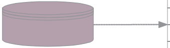
$$\begin{aligned} \underset{k, I_1, \ldots, I_k}{\text{minimize}} \quad & \sum_{i=1}^{k} |I_i| \\ \text{subject to} \quad & \sum_{i=1}^{k} \int_{I_i} f_X^{\mathcal{D}}(x)dx \geq \theta. \end{aligned}$$

Heterogeneity

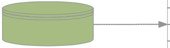
- Heterogeneity at three levels
  - Schema-level



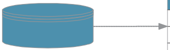
Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
...	...	...



City	Temp	Date
Burnaby	21	06/10/06
Vancouver	22	06/11/06
Richmond	18	06/12/06
Richmond	18	06/13/06
...	...	...



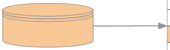
City	Temp	Date	...	Total Rain
Burnaby	19	10-June-06	...	0.2
Vancouver	17	11-June-06	...	0.0
Surrey	15	11-June-06	...	0.0
Vancouver	20	12-June-06	...	1.4
...	...	...	...	...



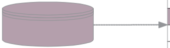
Location	Temp	Date	Total Snow	Total Rain
Surrey	15	06/11/06	0.0	0.0
Surrey	19	06/12/06	0.0	1.2
...	...	...	...	...

Heterogeneity

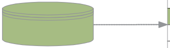
- Heterogeneity at three levels
  - Schema-level
  - Instance-level



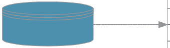
Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
...	...	...



City	Temp	Date
Burnaby	21	06/10/06
Vancouver	22	06/11/06
Richmond	18	06/12/06
Richmond	18	06/13/06
...	...	...



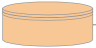
City	Temp	Date	...	Total Rain
Burnaby	19	10-June-06	...	0.2
Vancouver	17	11-June-06	...	0.0
Surrey	15	11-June-06	...	0.0
Vancouver	20	12-June-06	...	1.4
...	...	...	...	...



Location	Temp	Date	Total Snow	Total Rain
Surrey	15	06/11/06	0.0	0.0
Surrey	19	06/12/06	0.0	1.2
...	...	...	...	...

Heterogeneity

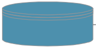
- Heterogeneity at three levels
  - Schema-level
  - Instance-level
  - Value-level



Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
...	...	...

City	Temp	Date
Burnaby	21	06/10/06
Vancouver	22	06/11/06
Richmond	18	06/12/06
Richmond	18	06/13/06
...	...	...

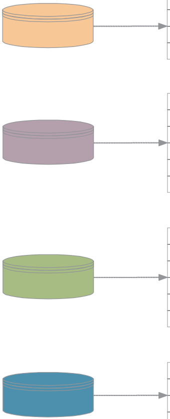
City	Temp	Date	...	Total Rain
Burnaby	19	10-June-06	...	0.2
Vancouver	17	11-June-06	...	0.0
Surrey	15	11-June-06	...	0.0
Vancouver	20	12-June-06	...	1.4
...	...	...	...	...



Location	Temp	Date	Total Snow	Total Rain
Surrey	15	06/11/06	0.0	0.0
Surrey	19	06/12/06	0.0	1.2
...	...	...	...	...

Value-level Heterogeneity

- Focus of our work is on value-level heterogeneity
  - Problem exists in various domains, e.g., stock, flight, weather domain
- Prior work assumes a single true answer exists, which we do not



Location	Avg Temp	Date
Burnaby	21	10-June-06
Vancouver	19	11-June-06
...	...	...

City	Temp	Date
Burnaby	21	06/10/06
Vancouver	22	06/11/06
Richmond	18	06/12/06
Richmond	18	06/13/06
...	...	...

City	Temp	Date	...	Total Rain
Burnaby	19	10-June-06	...	0.2
Vancouver	17	11-June-06	...	0.0
Surrey	15	11-June-06	...	0.0
Vancouver	20	12-June-06	...	1.4
...	...	...	...	...

Location	Temp	Date	Total Snow	Total Rain
Surrey	15	06/11/06	0.0	0.0
Surrey	19	06/12/06	0.0	1.2
...	...	...	...	...

Processing overhead of operations

- KDE dominates the processing overhead for extracting statistics
- Sampling the viable answers dominates the overall time needed for sampling and extracting statistics