EVALUATION OF GEOCODING SERVICES AND REFERENCE DATASETS FOR BRITISH COLUMBIA

VANCOUVER GIS USERS' GROUP

DECEMBER 9, 2015

SUNNY MAK AND TIFFANY KWONG

DISCLAIMER

- **We declare that we have no competing interests**
- **The findings and opinions expressed here are those of the authors and do not necessarily represent those of the British Columbia Centre for Disease Control**

ACCESS TO DATASETS

- **Via public access, subscription, provincial government and academic affiliation**

BACKGROUND

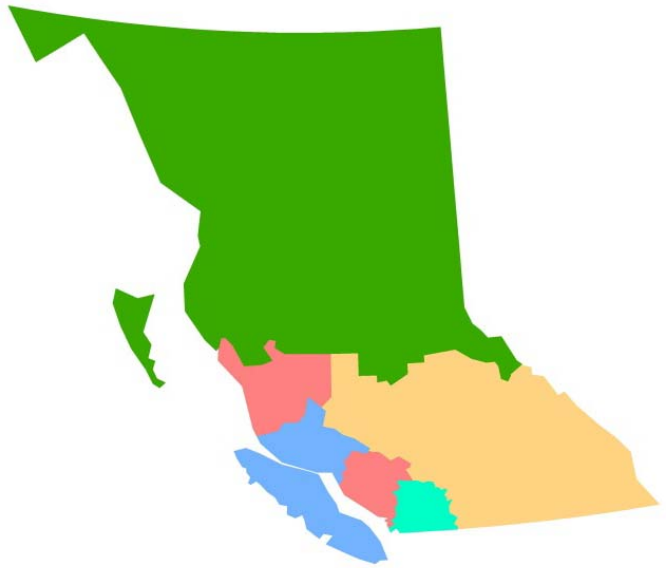
- **Geocoding is the process of converting addresses into latitude and longitude coordinates**
- **Multiple geocoding services and reference datasets exist**
- **Industry sponsored project for BCIT GIS program**
- **Evaluation performed January to May 2015**

PURPOSE

- **Evaluate and compare 6 geocoding services and reference datasets**
- **Stratify by Health Authority to assess regional variations**

METHODS AND DATA (1)

- **BC Assessment Fabric cadastres and AddressBC civic addresses**
    - X,Y coordinates treated as the ‘true’ location
- **200 randomly selected from each of the 5 regional health authorities**



METHODS AND DATA (2)

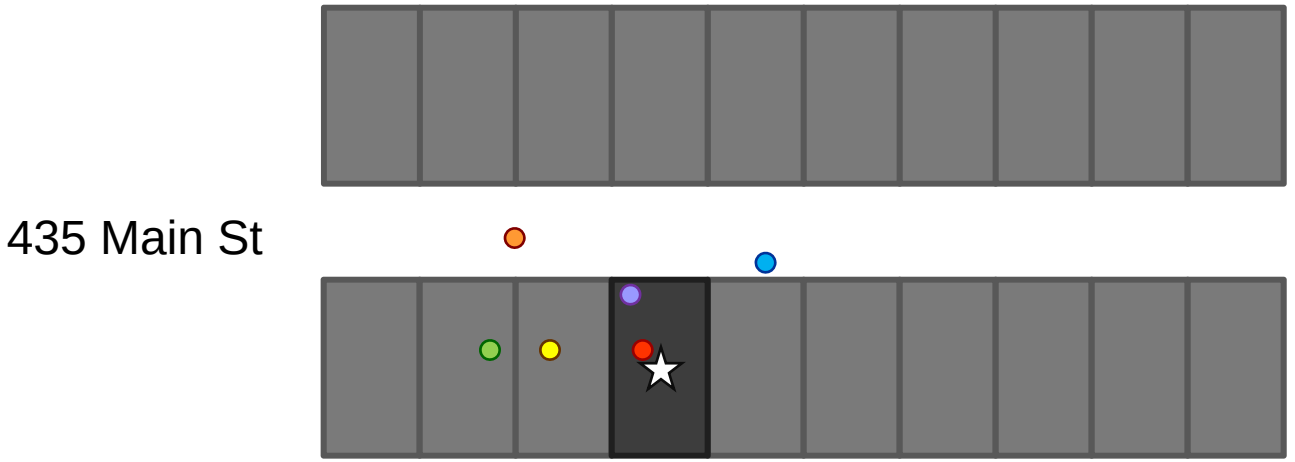
- **3 reference street network datasets**
  - Statistics Canada Road Network File (RNF)
  - dmtiSpatial CanMap Streetfiles
  - Province of BC Digital Road Atlas (DRA)
- **3 online geocoding services**
  - Google Earth Pro
  - dmtiSpatial Location Hub
  - Province of BC Physical Address Geocoder (PAG)

METHODS AND DATA (3)

- **ArcGIS for RNF, CanMap & DRA datasets**
  - US Addresses – Dual Ranges locator
    - From/To Left/Right
    - Prefix/Suffix Direction
    - Prefix/Suffix Type
    - Street Name
    - Left/Right City
  - Geocoding options
    - Minimum match score: 80
    - Side offset: 30 meters (determined with Near function)
  - Automatic and manual matching

METHODS AND DATA (4)

- **Metrics: geocoding success and positional accuracy**
  - Euclidian distance: $d(x,y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$



METHODS AND DATA (5)

- **Statistical analysis of positional accuracy**
  - Mean values are sensitive to outliers
  - Not normally distributed → non-parametric testing
  - Comparison of medians of more than 2 groups
    - Kruskal-Wallace H test
    - Mann-Whitney U test

RESULTS (1)

- **Match rate and positional accuracy**

Geocoding Method	Match Rate (%)	Minimum Dist. (m)	Mean Dist. (m)	Median Dist. (m)	Maximum Dist. (m)
RNF	71.7	7	83	39	3567
CanMap	87.8	7	89	44	3504
DRA	98.9	7	97	44	15347
Google	94.3	1	577	38	261358
PAG	99.3	0	81	37	4552
LocHub	95.5	0	1088	10	403865

RESULTS (2)

- **Statistical analysis of median distance**

Geocoding Method	RNF	CanMap	DRA	Google	PAG	Median Dist. (m)
RNF	-	-	-	-	-	39
CanMap	0.054	-	-	-	-	44
DRA	0.272	0.370	-	-	-	44
Google	0.037	*<0.001	*0.001	-	-	38
PAG	*<0.001	*<0.001	*<0.001	*<0.001	-	37
LocHub	*<0.001	*<0.001	*<0.001	*<0.001	*<0.001	10

*Statistical significance at 0.05 level with multiple testing correction (i.e. 0.05/15=0.0033)

RESULTS (3)

- **Regional variations**

Geocoding Method	Interior		Fraser		Vancouver Coastal		Vancouver Island		Northern	
	Match (n)	Median Dist. (m)	Match (n)	Median Dist. (m)	Match (n)	Median Dist. (m)	Match (n)	Median Dist. (m)	Match (n)	Median Dist. (m)
RNF	131	58	156	44	167	32	153	38	110	41
CanMap	162	60	189	47	197	38	181	41	149	44
DRA	195	56	199	45	200	31	200	45	195	58
Google	175	57	196	39	199	31	196	36	177	51
PAG	197	41	200	35	200	8	199	39	197	58
LocHub	185	17	198	14	200	4	193	6	179	17

200 randomly selected addresses from each Health Authority

DISCUSSION (1)

- **Among reference street network files:**
  - Digital Road Atlas has highest match rate (98.9%)
  - No statistical difference in positional accuracy among DRA, RNF and CanMap matches
- **Among online geocoding services:**
  - Physical Address Geocoder has highest match rate (99.5%)
  - Location Hub yields the most positionally accurate locations
    - However, errors can be large when it is wrong
  - Google Earth Pro does not report match method or score

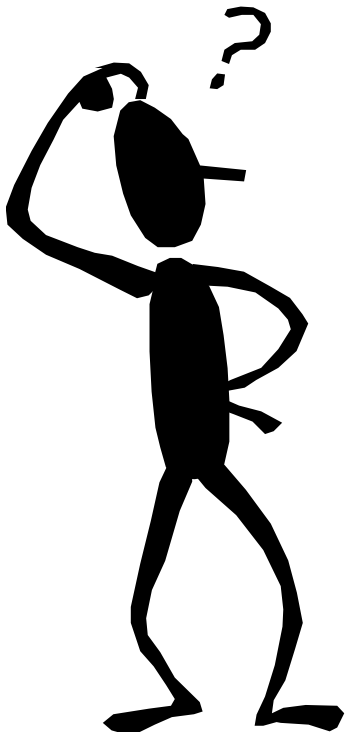
DISCUSSION (2)

- **Geocoding match success and positional accuracy higher in urban areas; lower in rural areas**
    - Highest in Vancouver Coastal region (84-100% match rate; 4-38m median distance)
    - Followed by Fraser region (78-100% MR; 14-47m MD) and Vancouver Island region (77-100% MR; 6-45m MD)
- **RNF had lowest geocoding match success**
    - Lowest number of street address range values in the attribute table

DISCUSSION (3)

- **Additional criteria for determining the 'best' geocoding service and/or reference dataset:**
  - Cost: free vs subscription/license for services/datasets
  - Privacy considerations: online vs offline
    - Location of residence is identifiable information
    - Protected health information cannot be disclosed to an external organization
  - Currentness: maintenance and update schedule
    - Historical versioning
  - Data preparation and standardization
    - Suite/unit numbers in addresses
    - Alias tables for alternate street and city names

QUESTIONS AND COMMENTS



sunny.mak@bccdc.ca

lai.tkwong@gmail.com