Using PCCF+ for coding and analysing health data: an introduction

Russell Wilkins
Health Information and Research Division
Statistics Canada, and Department of Epidemiology and Community Medicine, University of Ottawa

Health Users Group, SAS Institute, Toronto
Thursday 1 April 2010

A show of hands, please…

- How many are current users of PCCF+?
- Of those, how many are generally non-SAS users who run it as a black box?
- How many are currently not using PCCF+, but are considering it for future use?
- How many are here just for the other talks, and couldn’t care less about PCCF+?

Outline of today’s talk

- Introduction: possible uses and some examples
- Standard geographic variables and naming conventions
- How to use PCCF+
- Additional resources, limitations, etc

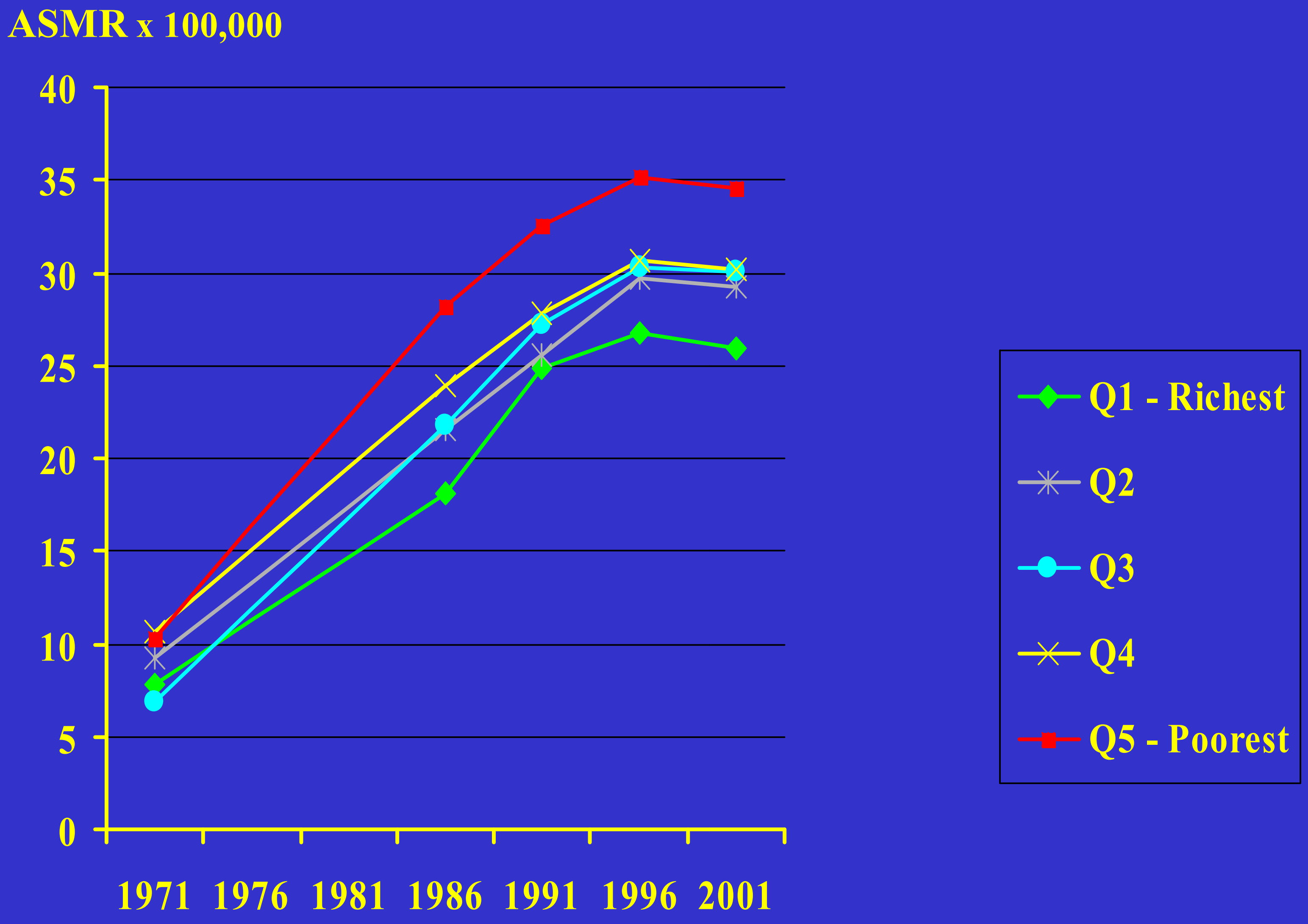
Possible uses of small-area data

- Neighbourhood SES (as determinant or confounder)
- Proxy or to help impute missing data for income, ethnicity
- Add policy relevance by aggregating to administrative areas, health planning units, school districts, etc.
- Deal with changes over time: newly created geographic units and revised boundaries (amalgamations, splits)
- Point-to-point distance, road distance, travel time
- Analysis by community characteristics
  - water supply, air pollution, UV radiation, social cohesion, access to services, parks, urban-rural-MIZ, segregation, etc.
- To permit studies of migration over time (for exposure or SES histories, or for better access to services, etc.) when longitudinal files are available
- Additional identifiers for record linkage purposes

Examples from earlier studies

- Lung cancer mortality trends among females, by neighbourhood income quintile, 1971-2001
- Probability of survival to age 75, by family vs neighbourhood income quintile, about 1996
- Distance to nearest school, and university participation
- Incident events mapped against environmental exposures
- Aboriginal-area life expectancy (geozones)*

Lung cancer mortality, females



Distance to post-secondary education

- Marc Frenette. *Too far to go on? Distance to school and university participation.* Research Paper Series, Analytical Studies No. 191. Ottawa: Statistics Canada catalogue 11F0019 No. 191, 2004.

- http://www.statcan.ca/english/research/11F0019MIE/11F0019MIE2002191.pdf

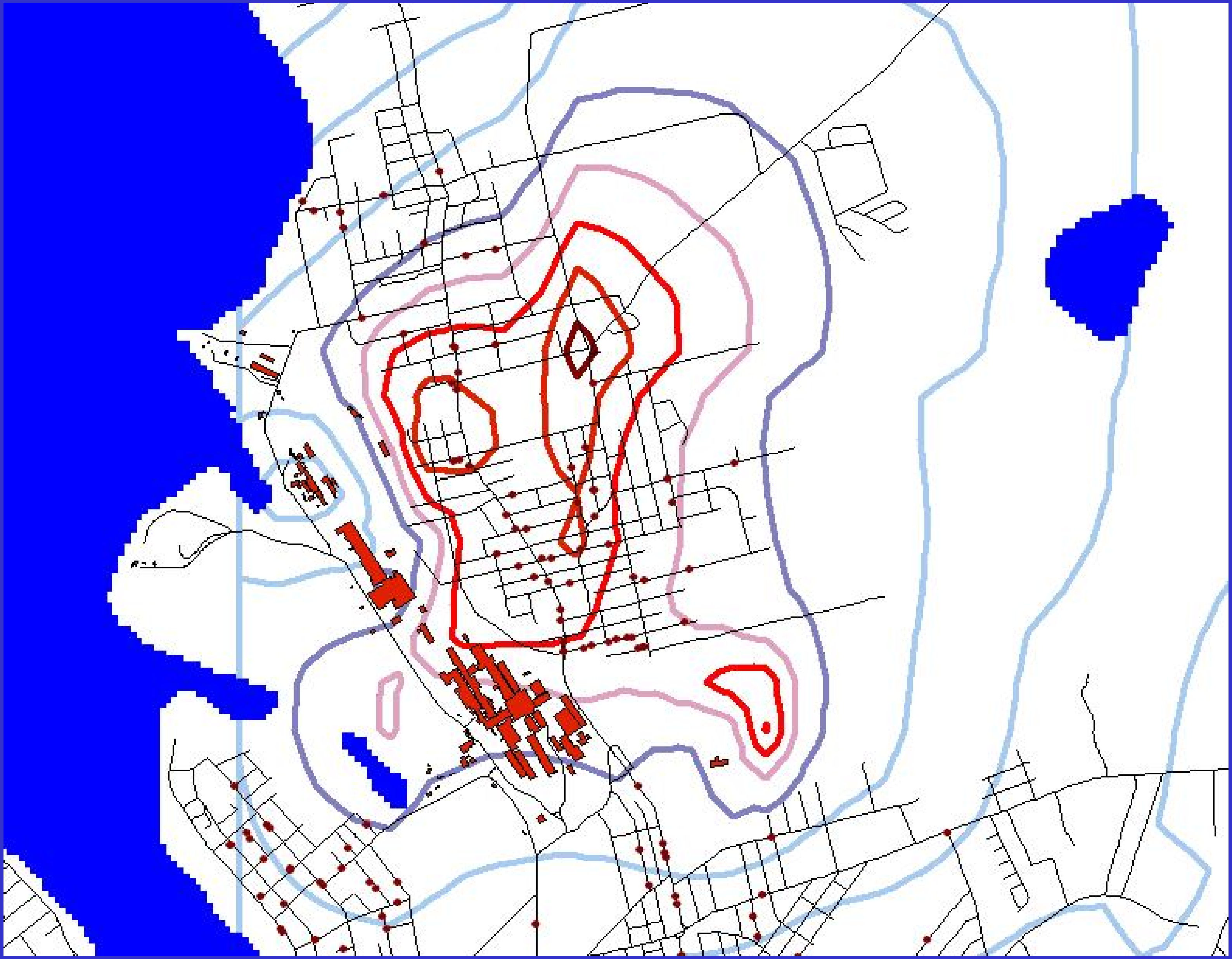
Data / Methods / Findings

- Survey of labour and income dynamics (SLID) 1993-1998 (postal codes while in high school); List of university postal codes; PCCF+

- After controlling for family income, parental education, and other factors associated with university participation, students living 'out-of-commuting distance' were far less likely to attend university than students living within commuting distance (<40 km). Dose-response by distance.

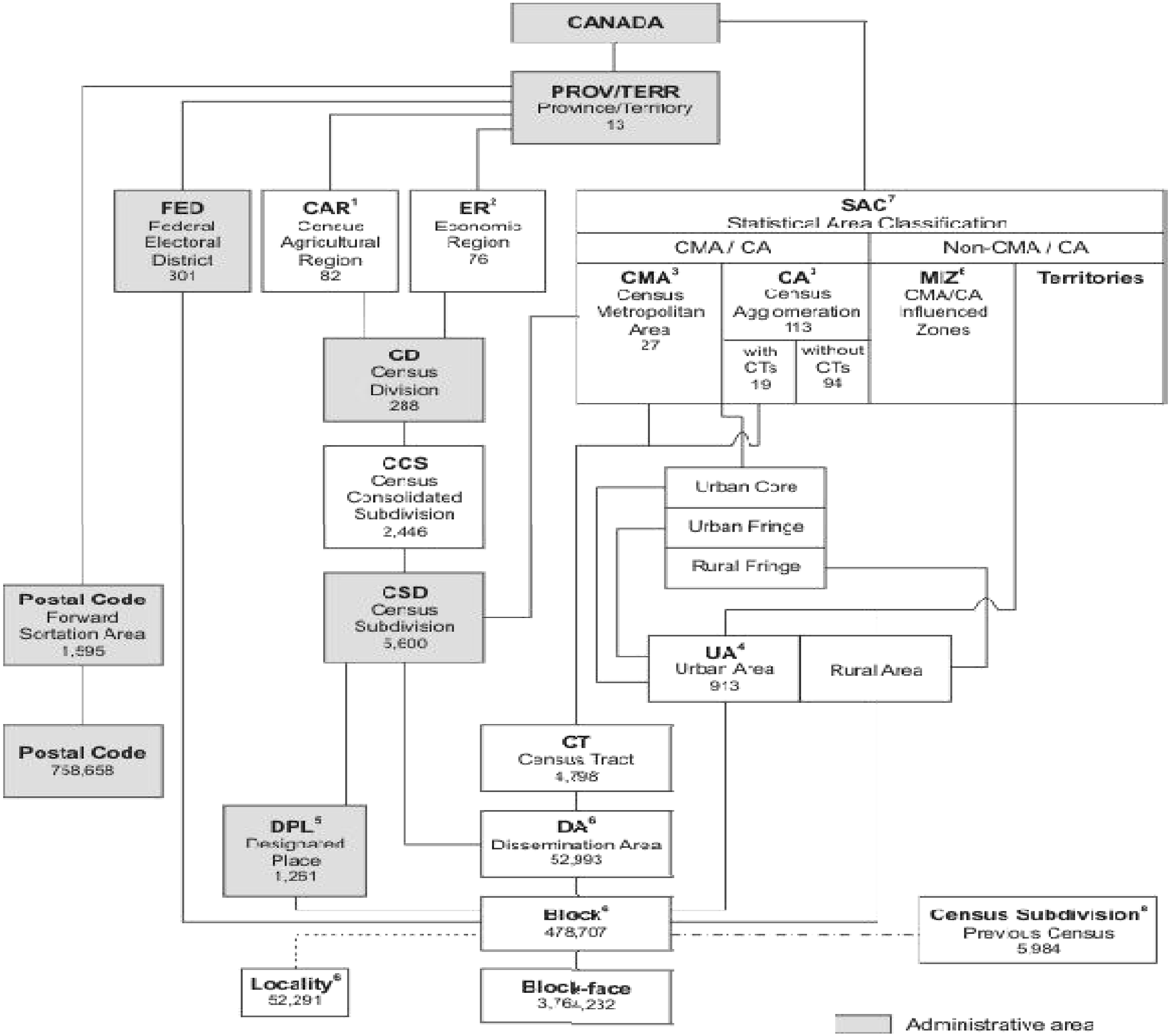
Sidney tar ponds environmental health study

- Geographic links directly from addresses, so increased resolution for a small urban area where block face coding not available on PCCF
- Illustrates GIS-based approach
- Events assigned to latitude and longitude
- Street network and pollution overlays
- Air photo and satellite images integrated



Census standard geography

- Lots of levels, most like Russian dolls
- Some levels defined analytically, others administratively
- Changes occur “only” every 5 years (even though administrative boundaries change continuously)



▨ Administrative area

□ Statistical area

Linkage using point-in-

[1] Census agricultural regions in Saskatchewan are composed of census consolidated subdivisions.

[2] Economic regions in Ontario are composed of municipalities (census subdivisions).

Unambiguous naming convention: geoYYuid

- geo => geographic level in census hierarchy
  - DA, CT, CSD, CMA, etc.
- YY => vintage of census geography required
  - DA01uid ≠ DA06uid (≈ 30% changed)
- uid => unique identifier
  - higher levels always needed with ‘geo’
  - DA=PR(2)+CD(2)+DA(4)=8 digits, not just last 4

Why PCCF+?
Canadian postal codes can be tricky

- Population weights
- Diagnostics
- Imputations
- Supplemental codes
- Reproducible, documented processing

Major problems which are dealt with by *PCCF*+

- Postal codes serving several DAs or blocks (especially in rural areas)
- Postal codes used by businesses or public institutions
- Postal codes which the regular PCCF only links to post office geography (rather than place of residence or business)
- Finding earlier “vintage” DA or EA, etc.

Black box: input => output

- Preparing directories and filenames
- Reading in the data to be coded
- Dealing with the problems identified (.PRB)
- Using the HLTHOUT file (.GEO)

Residential versus Institutional

- GEORES5F.SAS
  - Use to code records where the postal code is for a place of residence

- GEOINS5F.SAS
  - Use to code records where the postal code is for a health care facility, doctor's office or other institution or business

Summary of results

APPENDIX D:
SAMPLE OUTPUTS
FROM THE *PCCF+* PACKAGE

```
SUMMARY OF AUTOMATED CODING RESULTS USING GEOCODES/PCCF VERSION 5

RECORDS  PERCENT  PROB     MESSAGE          ACTION
-------------------------------------------------------------------------------------
   3996   100.00  TOTAL RECORDS INPUT FROM HLTHDAT (ID + PCODE)
    131     3.28  0 ERROR: NO MATCH TO PCCF---CHECK PCODE/ADDRESS &OR CODE MANUALLY
      5     0.13  1 ERROR: LINKED TO PO GEOG--CODE MANUALLY IF RESID ADD AVAILABLE
      3     0.08  2 WARNING: NON-RESIDENTIAL--CHECK PCODE/ADDRESS (LEGITIMATE RES?)
      3     0.08  3 WARNING: BUSINESS BLDG----CHECK PCODE/ADDRESS (LEGITIMATE RES?)
    241     6.03  4 WARNING: COMMERC/INSTITU--CHECK PCODE/ADDRESS (LEGITIMATE RES?)
     65     1.63  5 WARNING: RETIRED PCODE----CHECK PCODE/ADDRESS IF OLD DMT UNKNOWN
      1     0.03  6 NOTE: MULT MATCH CSD-PCCF-DISTRIBUTED AMONG APPLIC DA/BLK/BLKF
    535    13.39  7 NOTE: MULT MATCH CSD-WCF--DISTRIBUTED BY POP WEIGHTS OBSERVED
   3012    75.38  9 NO PROB (ERR,WARN,NOTE)---NO ACTION REQUIRED
-------------------------------------------------------------------------------------
      8     0.20  NOT CODED AT ALL
     39     0.98  PARTIALLY CODED TO PR ONLY
      2     0.05  PARTIALLY CODED TO PR + (CD OR CMA)--& APPROX LAT LONG
     12     0.30  PARTIALLY CODED TO PR+CD+CMA--AND APPROX LAT LONG
      8     0.20  PARTIALLY CODED TO PR+CD+CMA+CSD--AND APPROX LAT LONG
   3927    98.27  FULLY CODED TO PR+CD+CMA+CSD+CT+BLK--AND DA/BLK/BLKFACE LAT LONG
-------------------------------------------------------------------------------------
```

Coded output files (HLTHOUT+GEOPROB)

GEOG CODING

- ID (<=12), PCODE
- PR, CD, CSD
- CMA, CT; HR, SUB
- DA, BLK; DA06uid
- LAT, LONG
- QAIPPE, CSIZE, MIZ
- SACTYPE, NSREL
- RESFLG, INSTFLG
- EA81uid-EA96uid, DA01uid
- ER, AR, CCS, BLKURB, DPL

DIAGNOSTICS

- DMT, DMTDIFF
- LINK (PROB)
- SOURCE
- NCSD, NCD
- RPF, SERV, PREC
- BLDG NAME+ADR*
- CSDNAME+TYPE*
- CPCCODE
- RESFLG, INSTFLG

```
GEOCODES/PCCF VERSION 4 -- SAMPLE OUTPUT FROM THE HLTHOUT DATASET (.GEOG1 FILE)
----------------------------------------------------------------------------------------------------------------------------
ID          PCODE  PRCDCSD CMA CT     DABLK  LAT     LONG      DPL DIAG      VER COMM HRSUB C Q S N U FED ER AR CCS EA96UID
----------------------------------------------------------------------------------------------------------------------------
1304183010  H1A5H8 2466025 462 580.03 000601 45689925073486893 000 A9D111172 R4A 3276 06    1 3 1 S 1 044 40 06 025 24045417
1304183033  H1A5G4 2466025 462 582.01 292702 45653189073503887 000 A9D111176 R4A 3276 06    1 3 1 S 1 044 40 06 025 24045358
1304183332  G1H2C1 2423030 421 273.01 082102 46856140071245151 000 A9D11117. R4A 2587 03    2 2 1 S 1 015 20 03 030 24016455
1304183333  G1H7B3 2423030 421 273.01 081902 46850294071240870 000 A9F111191 R4A 2587 03    2 2 1 S 1 015 20 03 030 24016452
1304183632  G8T8L9 2437055 442 200.00 015910 46367087072500828 000 B9D111171 R4A 2561 04    3 1 1 S 1 014 70 04 050 24014354
1304184533  J8V2P3 2481015 505 841.03 037904 45515264075736270 000 A9D111176 R4A 2752 07    2 3 1 S 0 023 60 08 015 24015556
1304185031  G1P1H6 2423025 421 039.02 065901 46822089071329615 000 A9D11117. R4A 3313 03    2 1 1 S 1 052 20 03 025 24054103
1304185033  G2E5Y7 2423055 421 140.03 047503 46806119071370503 000 A9D111173 R4A 2859 03    2 4 1 S 1 052 20 03 060 24054063
1601001210  L1G3Y1 3518013 532 015.00 008602 43937498078876105 000 A9D11117. R4A 5227 0330  3 2 1 S 1 016 30 03 013 35016270
1601002733  L8V3V5 3525005 537 005.01 059702 43217763079851251 000 A9F111191 R4A 4809 0837  2 1 1 S 1 030 50 01 005 35030108
1601005410  R2G0E6 4611040 602 141.02 071402 49937939097087637 000 A9D11117. R4A 6221 10    2 2 1 S 1 013 50 09 040 46008417
1601007832  P7A5G4 3558004 595 015.00 014505 48438993089226888 000 A9F111191 R4A 5549 1662  3 1 1 S 1 087 95 05 004 35084320
1601007833  P7B3H1 3558004 595 011.01 031611 48421824089235996 000 A9F111191 R4A 5549 1662  3 1 1 S 1 087 95 05 004 35084410
1601009010  M6S4Y8 3520005 535 050.01 147401 43637293079471415 000 B9F111191 R4A 5562 0495B 1 4 1 S 1 064 30 03 005 35063258
1601009033  M6P2H9 3520005 535 100.00 140201 43664058079462540 000 A9F111191 R4A 5562 0495E 1 3 1 S 1 064 30 03 005 35098002
1601010231  K7M7B4 3510010 521 014.00 013602 44250712076533691 000 B9D111171 R4A 4951 0241  3 1 1 S 1 036 15 04 010 35037506
1601011533  L5C3S8 3521005 535 527.08 069101 43577841079654532 000 A9D111172 R4A 5106 0653  1 3 1 S 1 046 30 02 005 35049404
1601011910  S0E1E0 4714076 000 000.00 002410 53349268104019508 000 W7C934459 R4A 6735 08    5 1 0 R 1 006 50 8A 072 47002573
1601013832  L7R4M7 3524002 537 207.01 053802 43334767079821521 000 B9F111191 R4A 4458 0636  2 3 1 S 1 010 50 02 002 35008115
1601014733  L2G3E7 3526043 539 203.01 006904 43070976079095668 000 A9F111191 R4A 5177 0946  3 2 1 S 1 052 50 01 043 35051016
1601015931  L4W1L1 3521005 535 527.05 032501 43624059079608402 000 A9F111191 R4A 5106 0653  1 1 1 S 1 047 30 02 005 35047351
1601016133  L2S2M9 3526053 539 003.01 037804 43145861079253296 000 A9F111191 R4A 5473 0946  3 1 1 S 1 051 50 01 053 35090216
1601017132  L4N2V4 3543042 568 005.00 038106 44367352079679190 000 A9F111191 R4A 4358 0560  3 5 2 S 1 002 40 02 042 35079159
1601017421  N7S5L7 3538030 562 102.02 015804 42973744082365802 000 A9F111191 R4A 5391 1242  4 3 2 S 1 071 70 01 030 35072209
1601017633  M4K1C1 3520005 535 069.00 383001 43669948079342406 000 A9F111191 R4A 5562 0495I 1 2 1 S 1 008 30 03 005 35006061
1601017910  N4B2W4 3528052 547 000.00 008009 42756837080558774 000 H9C114259 R4A 4613 1034  4 4 3 S 0 027 50 01 052 35018012
1601018131  N6G2E5 3539036 555 044.04 035003 43006922081306309 000 A9D11117. R4A 5013 1144  3 3 1 S 1 044 60 01 036 35045463
1601019332  L5G1J8 3521005 535 540.01 037901 43553413079585884 000 B9F111191 R4A 5106 0653  1 1 1 S 1 048 30 02 005 35048068
1601019721  R2K0V9 4611040 602 133.00 070502 49927590097100976 000 A9F111191 R4A 6221 10    2 2 1 S 1 014 50 09 040 46014203
1601020010  M4E3M6 3520005 535 022.00 379901 43677506079285931 000 A9D11117. R4A 5562 0495K 1 5 1 S 1 003 30 03 005 35002068
1601020131  T7P1A3 4813031 000 000.00 004620 54164822113845804 000 A9F112181 R4A 7709 26    5 4 0 R 1 001 70 06 028 48001057
1601020432  N4G4T7 3532004 546 000.00 007010 42876846080729595 000 B9F112181 R4A 5555 1152  4 4 3 S 1 063 60 01 012 35062064
1601020610  M1C1K9 3520005 535 362.02 374802 43788038079163502 000 A9D11117. R4A 5400 0495M 1 5 1 S 1 075 30 03 005 35077052
1601025533  T5H2X1 4811061 835 046.00 020303 53550678113501115 000 A9F111191 R4A 7229 25    2 1 1 R 1 015 60 05 061 48012253
1601026631  K1V9K4 3506008 505 002.05 087501 45347074075665245 000 B9F111191 R4A 5230 0151  2 3 1 S 1 060 10 04 008 35059014
1601027832  S4V0G7 4706027 705 008.02 019701 50432251104564832 000 A9D11117. R4A 6814 04    3 5 1 S 1 013 10 2B 027 47007161
1601028831  N7S4X8 3538030 562 102.02 015903 42970869082365165 000 A9F111191 R4A 5391 1242  4 2 2 S 1 071 70 01 030 35072208
1601028832  N7T6J8 3538030 562 008.00 019504 42982172082396827 000 A9F111191 R4A 5391 1242  4 2 2 S 1 071 70 01 030 35072164
1601029531  T1K4A4 4802012 810 019.00 016101 49678240112881944 000 A9D11117. R4A 7414 20    4 2 2 S 1 018 10 02 011 48017419
1601030710  L5C3L4 3521005 535 527.08 069502 43576525079661365 000 A9F111191 R4A 5106 0653  1 4 1 S 1 046 30 02 005 35049405
1601030733  L5A3T1 3521005 535 521.06 085901 43597525079626646 000 B9F111191 R4A 5106 0653  1 2 1 S 1 047 30 02 005 35047113
1601031231  L8N2Z3 3525005 537 033.00 044701 43246956079851089 000 A9F111191 R4A 4809 0837  2 1 1 S 1 029 50 01 005 35032002
1601032031  K8A7W4 3547064 515 000.00 004912 45817759077093184 000 A9F112181 R4A 5256 0157  4 5 3 S 1 070 15 04 075 35068254
1601033332  R2K0K5 4611040 602 134.00 071204 49930495097093590 000 A9F111191 R4A 6221 10    2 3 1 S 1 014 50 09 040 46014208
1601035633  R2C5B2 4611040 602 120.02 085503 49900542096969280 000 A9F111191 R4A 6221 10    2 4 1 S 1 014 50 09 040 46014003
----------------------------------------------------------------------------------------------------------------------------
```

The problem file (.PRB)

- Unmatched to any known postal code
- Matched but only linked to PO geography
- Non-residential postal codes
- Postal codes usually for business buildings
- Postal codes for commercial / institutional buildings – check if legitimate residence

```
Sample printout from the GEOPROB dataset       GEOCODES/PCCF VERSION 4
                                          PARTIAL PRINT OF GEOPROB FILE (ERRORS & WARNINGS, BUT NO NOTES)
ID          PCODE  PRCDCSD CMA CT     DABLK  LL   HRSUB DPL DIAG       BLDG NAME,ADR(CPCOMM:CMA/DPL) :CDNAME        CDTYP CSDNAME TY
------------------------------------------------------------------------------------------------------------------------------------
0 ERROR: NO MATCH TO PCCF---CHECK PCODE/ADDRESS &OR CODE MANUALLY
------------------------------------------------------------------
1202050810  A1X5J7 1001485 001 301.02 013501 4705 01     000  90I31994. St. John's CMA                :Avalon Peninsul DIV CONCEPTIT*
1201026310  B2M5B3 1200999 999 999.99 999900 4506 99     999  902..892.                               :                             *
1302025710  G0K2K0 2410005 000 000.00 007009 4806 01     000  90I949949 NOT CMACA                     :Rimouski-Neiget MRC ESPRIT-SM*
1301031010  H9G3X9 2466140 462 521.01 235801 4507 06     000  90I31994. Montréal CMA                  :Montréal        CU  DOLLARD-V*
1602451310  K7K2T0 3510010 521 008.00 018405 4407 0241   000  90I11994. Kingston CMA                  :Frontenac       CTY KINGSTONC*
1604153110  M3Y4A1 3520005 535 999.99 999900 4307 99999 999  902..892. Toronto CMA                   :Toronto         DIV TORONTO C*
1604305110  R3N3L2 4611040 602 008.00 038001 4909 10     000  90I11994. Winnipeg CMA                  :Winnipeg        DIV WINNIPEGC*
1802106710  V1S4X1 5933042 925 006.00 004302 5012 14     000  90I21994. Kamloops CA1                  :Thompson-Nicola RD  KAMLOOPSC*
1802068310  V4T4J5 5935027 915 102.02 015502 4911 13     175  90I41994. Kelowna CA1:Westbank (UNP)    :Central Okanaga RD  CENTRAL RD
1803049810  V9C5T3 5917044 935 154.02 048004 4812 41     000  90I51994. Victoria CMA                  :Capital         RD  LANGFORDDM
------------------------------------------------------------------
1 ERROR: LINKED TO PO GEOG--CODE MANUALLY IF RESID ADD AVAILABLE
------------------------------------------------------------------
1604055531  R4J1A1 4611999 602 999.99 999900 4909 99     000 JZ1I22824. HEADINGLEY:Winnipeg CMA       :Winnipeg        DIV           *
1201059710  A1X4G9 1001999 001 999.99 999900 4705 99     000  K1I318341 BOX 18001:18060 STN MAIN UPPER GULLIES                      *
------------------------------------------------------------------
2 WARNING: NON-RESIDENTIAL PCODE--CHECK PCODE/ADDRESS (LEGIT RES?)
------------------------------------------------------------------
1304154932  H3L1B9-2400999 462 999.99 999900  . . 99     999  E2F119191 CENTRE MEDICAL HENRI-BOURASSA 222 HENRI-BOURA MONT          *
1603422510  L4C9S7-3500999 535 999.99 999900  . . 99999 999  E2F119191 BUSINESS BUILDING 120 NEWKIRK RD RICHMOND HILL              *
1602226510  T2S2T6-4800999 825 999.99 999900  . . 99     999  E2F119191 FOODVALE OFFICE COMPLEX 5005 ELBOW DR SW CALGARY            *
1601088310  T5N4A3-4800999 835 999.99 999900  . . 99     999  E2F119191 PEOPLES TRUST PLAZA 10216 124 ST NW EDMONTON                *
1302161110  H3N2Y1-2400999 462 999.99 999900  . . 99     999  G2F119191 VIDEOTRON LTEE 405 OGILVY AV 200 MONTREAL                   *
1804030033  V2A5A9-5900999 913 000.00 999900  . . 99     999  G2D119171 CITY OF PENTICTON 171 MAIN ST PENTICTON                     *
------------------------------------------------------------------
3 WARNING: BUSINESS BLDG----CHECK PCODE/ADDRESS (LEGITIMATE RES?)
------------------------------------------------------------------
1604118533  L6Y2N4@3521010 535 572.05 020201 4307 0653   000  E3F111191 APARTMENT BLDG 430 MCMURCHY AVE S BRAMPTON         BRAMPTONC*
1604503732  T5H4B9@4811061 835 046.00 020808 5311 25     000  E3F111191 HYS MEDICAL CENTRE 11010 101 ST NW EDMONTON        EDMONTONC*
------------------------------------------------------------------
4 WARNING: COMMERC/INSTITU--CHECK PCODE/ADDRESS (LEGITIMATE RES?)
------------------------------------------------------------------
1801082533  V5G4J3?5915025 933 230.01 139201 4912 22     000 BG4F111191 BRITISH COLUMBIA INSTITUTE OF TECHNOLOGY 4200 BURN BURNABY C*
1202190833  A1B1S5@1001519 001 013.00 025301 4705 01     000  G4F111191 ST PATRICKS MERCY HOME 146 ELIZABETH AVE ST. JOHN' ST. JOHNC*
1202154133  A2A2E1@1006017 010 000.00 003010 4805 03     000  G4D112171 CENTRAL NEWFOUNDLAND REGIONAL HEALTH CENTRE 5 GRAN GRAND FAT*
1303089633  H2C3H6@2466025 462 277.00 265801 4507 06     000  G4F111191 LES RESIDENCES LAURENDEAU,LEGARE,LOUVAIN 1725 MONT MONTRÉALV*
1603169333  M1H3A1@3520005 535 356.00 361001 4307 0495N 000  G4F111191 CEDARBROOK LODGE 520 MARKHAM RD SCARBOROUGH        TORONTO C*
1602154410  M9W4L3@3520005 535 246.00 184101 4307 0495A 000  G4F111191 KIPLING ACRES HOME FOR THE AGED 2233 KIPLING ETOBI TORONTO C*
1604515931  N2L3G1@3530016 541 106.01 029605 4308 0765   000  G4F111191 UNIVERSITY OF WATERLOO 200 UNIVERSITY AVE W WATERL WATERLOOC*
1604443433  R1N3V4@4609029 607 000.00 001414H4909 40     000  G4F112181 LION'S PRAIRIE MANOR 24 9TH ST SE PORTAGE LA PRAIR PORTAGE C*
1603468632  R3N1V9@4611040 602 510.02 036601 4909 10     000  G4F111191 CANADIAN FORCES BASE WINNIPEG, KAPYONG BARRAC WINN WINNIPEGC*
1601086332  R7N1R7@4617050 000 000.00 001114 5110 60     000  G4F111191 DAUPHIN GENERAL HOSPITAL 625 3RD ST SW DAUPHIN     DAUPHIN C*
1603548732  S4S3B4@4706027 705 002.02 049002 5010 04     000  G4F111191 EXTENDICARE/PARKSIDE 4540 RAE ST REGINA            REGINA  C*
1602539533  T5K0L4@4811061 835 032.02 015604H5311 25     000  G4F111191 GENERAL HOSPITAL 11111 JASPER AVE NW EDMONTON      EDMONTONC*
1803100131  V6T1K2@5915020 933 069.00 094705 4912 32     000  G4D111171 WALTER GAGE RESIDENCE ( UBC ) 5959 STUDENT UN VANC GREATER RD
-----------------------------------------------------------------------------------------------------------------------------------
```

Code your data only once, but analyse them many times

- Be sure to correct all serious problems identified by the automated coding. It usually takes a couple of iterations to get the whole file clean.
- The importance of the problems identified by the diagnostic codes depends on the data set and on the analyses to be done. Retain the diagnostic codes!
- Once coded, the same dataset can be used for various kinds of studies (eg SES disparities, access to services, environmental health).

What problems have you encountered using PCCF+?

- Virtually all the “features” of PCCF+ are the result of fixes to former problems identified by users.
- Examples: flagging of non-residential postal codes; look up of building names and addresses; population-weighted assignments; imputations (now at 3, 4, and 5 digits); earlier vintage codes.

User input needed

- Reporting errors encountered
  - Entire streets assigned to single urban pcode
  - WCF can easily be edited
- Info for updating the EGMRES file
  - Easily updated as buildings classified
- Suggesting ideas for improvements
  - Need to impute for small EAs and DAs
  - Distances, historic geographies, sub-regional

Documentation

- Wilkins R. *PCCF+ Version 5F User's Guide*. Statistics Canada, 2010.

Getting help

- Talk to an experienced user
- Consult the documentation
- If that doesn't help, call Russell

Geographic tools / technical references

- Wilkins R. *PCCF+ Version 5F User's Guide.* Statistics Canada, 2010.
- Gonthier et al, Merging area-level census data with survey data in STC RDCs. *ITB: the Research Data Centres Information and Technical Bulletin* (12-002), 2006
- Wilkins R. Neighbourhood income quintiles derived from Canadian postal codes are apt to be misclassified in rural but not urban areas. HAMG internal report, 2004.

Concluding remarks

- Small area geography and/or latitude-longitude coordinates are increasingly becoming a part of most health data sets and are useful to at least some extent in most health studies, even where individual measures of SES are available.
- Familiarity with the methods (tools and techniques), as well as the strengths and limitations, of dealing with such data, will allow health researchers to meaningfully exploit their potential.
- But like with other methods, it’s not enough to just do it mechanically. Think through what you’re doing and why.

PCCF+ contacts:
Russell Wilkins & Saeeda Khan

Health Analysis Division
Statistics Canada, RHC-24
100 Tunney's Pasture Driveway
Ottawa ON   K1A OT6
Tel: 1-613-951-5305 (Russell) 951-4765 (Saeeda)
Fax: 1-613-951-3959
Email: russell.wilkins@statcan.gc.ca
Email: saeeda.khan@statcan.gc.ca

Saeeda Khan

- McGill health geography (with Nancy Ross); several years at STC/HAD
- Working with Eric Hortop (Methodologist, HSMD) re construction, updates and documentation of PCCF+
- "Passing the torch" after 2011 rebuild

Use of SLI for residential coding introduces systematic bias

- *Most* DAs in rural postal coded areas can never be coded
- *Many* CSDs in rural areas can never be coded
- *A high proportion* of the population in rural areas will be systematically miscoded (to wherever the SLI is situated)

Implications of such systematic biases introduced by use of SLI

- Serious numerator-denominator mismatch whenever census population (denominator) data are required
- “Hot spots” surrounded by “cold spots”
- Over-coding of UARA classification of “urban” (BLKURB, based on block-level density in rural village centres)

When is forced 1:1 coding from postal codes acceptable?

- For distance calculations, where all you really need is a single representative *average* location in the service area of the postal code.
- For calculations of rates based on denominators derived from the *same* file as the numerators, so that the coding errors will be in balance (systematically biased by the same amount in both the numerator and denominator). Example: for birth outcomes other than fertility rates.
- For calculation of rates based on denominators derived from another postal coded file *which was processed in the same way*, such as a provincial health insurance master beneficiary file.
- But you always need to check for non-residential (business-only) postal codes, and perhaps impute for partially incorrect codes, etc.

Misclassification

- In rural areas (and urban fringe) only, DA is assigned probabilistically—leading to random misclassification of DA and associated neighbourhood income quintile (QAIPPE).
- => reduced ability to detect effects in rural areas (lower RRs, RDs), but almost no impact in urban areas
- So be very careful in interpreting the expected lower effect estimates for rural vs urban areas. Such results may disagree with individual measures of SES.
- Working paper showing extent of misclassification and impact on RRs, plus correction factors which could be applied to help compensate for the misclassification.

Misclassification of QAIPPE?

Reference

- **Wilkins R. *Neighbourhood income quintiles derived from Canadian postal codes are apt to be misclassified in rural but not urban areas.* Health Analysis and Measurement Group, Statistics Canada, 2004-08-25. [Draft]**

Misclassification of income quintile in rural areas

- Neighbourhood income quintiles derived from Canadian postal codes are apt to be misclassified in rural but not urban areas.
- The extent of the misclassification has been evaluated, and a method of correction developed.
- The correction is of little effect in urban areas, but of considerable effect in rural areas.

- Wilkins R. HAMG working paper, 2004-08-25 Draft.

Pitfalls of automated coding: some examples (1)

- **Problem:** In a study of psychiatric problems among Manitoba children, dozens of children had the same downtown Winnipeg postal code.
- **Diagnosis:** Examination of the building name and address showed the postal code referred to the office of the provincial trustee responsible for minor children in provincial care. Use of the geography and neighbourhood characteristics associated with that postal code would have seriously biased the study results.
- **Solution:** Most non-residential postal codes including those for government and institutions can be identified by looking at the building / organization name and address in the problem output. Then either find the postal code for the true place of residence (if appropriate re study aims) or set geography to missing (as was done for this study).

Pitfalls of automated coding (2)

- **Problem:** In a study Quebec births, many births were for mothers with the same few urban postal codes. The delivery mode type of those postal codes was not B (for large apartment buildings).
- **Diagnosis:** It was determined that missing postal codes were being administratively assigned the postal code of the hospital of birth, so that health region could be assigned even though the mother's postal code was unknown. Use of the associated small-area geography and/or neighbourhood characteristics would have systematically biased the results.
- **Solution:** Identify postal codes for hospitals, which should not be accepted as place of residence of the mother. Then either use the address information (if available) to find the mother's own postal code or set geography to missing (as was done for this study).

Pitfalls of automated coding (3)

- **Problem:** In an early study using BC vital statistics data with nearly 100% presence of full postal codes, we were coding many deaths as residents of Montreal, Quebec, although the decedents had been born in other provinces or countries, and the provincial municipal coding showed BC place of residence.
- **Diagnosis:** The non-existent postal code H0H0H0 (ho-ho-ho!) was being assigned when no postal code was reported. PCCF+ imputed geography from partial postal codes, although error codes were also assigned.
- **Solution:** The full address was used to find a real postal code, or to assign geography manually if no postal code could be found.

Pitfalls of automated coding (4)

- **Problem:** The usual place of residence on vital statistics mortality files may legitimately include institutional addresses. How can we know when that is the case?
- **Diagnosis:** Systematically identify such cases by postal code (where unique) and by postal code and address (when not unique). In our studies of mortality by income, up to 15% of deaths are typically for residents of chronic care hospitals and other long-term health care facilities.
- **Solution:** Remove institutional residents from both deaths and population at risk (numerator and denominator). More of a problem for hospital separation data.

Pitfalls of automated coding (5)

- **Problem:** In a study set in the Kingston area, many health events were for a relatively few postal codes, which were not known to be hospitals or long-term health care facilities.
- **Diagnosis:** Closer examination showed them to be for prisons and university residences.
- **Solution:** Systematically identify such cases, and depending on the purposes of the study, decide whether or not to use such cases in the analysis. (Note: The smaller the study area, the greater the potential impact of such problems.)

Pitfalls of automated coding (6)

- **Problem:** In various studies, postal codes for businesses keep appearing in the field for place of residence, apparently not due to keying errors.
- **Diagnosis:** Likely a small but non-negligible proportion of persons either prefer to receive correspondence at their place of work, or mistakenly report the wrong postal code.
- **Solution:** Systematically identify postal codes for non-residential addresses. Try to recode based on street address or postal code reported on other records for the same person.

Pitfalls of automated coding (7)

- **Problem:** In a Nova Scotia study of socio-economic differentials in mental health based on person-oriented hospital data, the neighbourhood SES of the mentally ill, as determined from their current postal code, tended to decline over time.
- **Diagnosis:** Use of current postal code to assign neighbourhood SES would risk confusing cause with effect.
- **Solution:** In person-oriented analysis, assign neighbourhood SES based on postal code at *initial* hospitalization or diagnosis.

Pitfalls of automated coding (8)

- **Problem:** Some studies require geographic coding of business and industrial locations, including mines, manufacturing establishments and dumpsites.
- **Diagnosis:** The locations of such sites could be anywhere in the service area of postal code, unrelated to population distribution.
- **Solution:** The population-based assumptions on which resolution of multiple matches are made using PCCF+ are simply *not appropriate* for coding in such cases. Consider alternate coding methods based on nearest road intersection, retrieval of latitude and longitude information from other files, or use of GPS.