

**More about PCCF+ etc.
(for the hard core)**

**Russell Wilkins
Health Analysis and Measurement Group
Statistics Canada**

**Public Health Agency of Canada,
Thursday 27 January 2005**

Plan

- **Technical facts about PCCF+**
- **Pre-editing the postal code field**
- **Using the problem file listing**
- **Understanding the diagnostic fields produced**
- **Explanation of the supplementary geographic codes**
- **Translating across vintages of census geography**
- **Many-to-many distance calculations**
- **Evaluating coding accuracy**

Did you realize that ..

- PCCF+ is completely “open source”, and the SAS programs are highly commented to explain what’s happening at each step. For example, you could easily edit the programs to reduce unneeded print output listings.
- All files except those with .pdf and .doc extensions are ASCII plain text, which can be read with any text editor or word processor (such as Notepad, Word, or the SAS editor).
- The record layout of each of the component files is shown in the DATA step INPUT statement, so you could reuse those files for other work if you wished.
- Except for population-weighted random assignments to DA and BLK (which is tricky), only simple coding is used—mostly just SORT BY and MERGE BY, so you should be able to figure out what’s happening at any point, should you want to or need to do so.
- The distance calculation code (in GEORES4x.SAS or DIST4x.SAS) can be copied and reused for other applications. No need to re-program the spherical geometry calculations.

Pre-edits of postal code field

- All caps: `PCODE=UPCASE(PCODE);`
- Change O to 0 and change I or l to 1 in 2nd, 4th and 6th position
- Remove extraneous hyphens or blanks
- Ensure PCODE field is always same length
Consider replacing each missing character with a period (“.”)
- List all records with “illegal” form (see top of p. 17 of the *User’s Guide*)

Pre-edits of postal code field: listing all records with “illegal” form

- P1=SUBSTR(PCODE,1,1); P2=SUBSTR(PCODE,2,1); P3=SUBSTR(PCODE,3,1);
P4=SUBSTR(PCODE,4,1); P5=SUBSTR(PCODE,5,1); P6=SUBSTR(PCODE,6,1);
- IF P1 IN('A' 'B' 'C' 'E' 'G' 'H' 'J' 'K' 'L' 'M' 'N' 'P' 'R' 'S' 'T'
'V' 'X' 'Y') THEN P1OK=1; /* NOT D F I O U W Z */
- IF P3 IN('A' 'B' 'C' 'E' 'G' 'H' 'J' 'K' 'L' 'M' 'N' 'P' 'R' 'S' 'T'
'V' 'W' 'X' 'Y' 'Z') THEN P3OK=1; /* NOT D F I O U */
- IF P5 IN('A' 'B' 'C' 'E' 'G' 'H' 'J' 'K' 'L' 'M' 'N' 'P' 'R' 'S' 'T'
'V' 'W' 'X' 'Y' 'Z') THEN P5OK=1; /* NOT D F I O U */
- IF P2 IN ('0' '1' '2' '3' '4' '5' '6' '7' '8' '9') THEN P2OK=1;
- IF P4 IN ('0' '1' '2' '3' '4' '5' '6' '7' '8' '9') THEN P4OK=1;
- IF P6 IN ('0' '1' '2' '3' '4' '5' '6' '7' '8' '9') THEN P6OK=1;
- IF (P1OK=1 AND P2OK=1 AND P3OK=1 AND P4OK=1 AND P5OK=1 AND P6OK=1)
THEN PCODEOK=1;ELSE PCODEOK=0;
- PROC PRINT;WHERE PCODEOK=0;VAR ID PCODE;TITLE 'ILLEGAL FORM';RUN;

Working with the problem file listing

- Scan each section—you should be able to understand what’s “wrong” with each record.
- More serious problems come first, least serious problems last.
- Similar problems are grouped together.
- You can re-sort within the groups, and/or cut and paste to regroup by decisions.
- Decide what to do re each case or category. Use the diagnostic codes to help.

Understanding the diagnostic codes (on both files)

- **LINK (PROB)**
- **DMT, DMTDIFF**
- **SOURCE**
- **PREC**
- **RPF, SERV**
- **NADR; NCD*, NCSD*** (*HLTHOUT only)

LINK (PROB)

- **0 Error** No match to PCCF
- **1 Error** Linked to PO geography
- **2 Warning** Non residential
- **3 Warning** Business building (usually)
- **4 Warning** Commercial / Institutional
- **5 Warning** Retired postal code, former DMT unknown
- **6 Note** Multiple CSD--by PCCF dup
- **7 Note** Multiple CSD--by WCF
- **9 No prob** No problem (error, warn, note)

DMT: Services to urban vs rural areas and business vs residential uses

- *For urban areas:*
- DMT=A B E G M
- 21 million persons
- 72% of population
- 45 persons/cen pcode
- 1.3 PCCF records
/pcode (1 EA/rec)
- EGM often business

- *For rural areas:*
- DMT=W H J K T X Z*
- 8 million persons
- 28 % of population
- 1200 pers/cen pcode
- 5 PCCF records/pcode
(span multiple EAs)
- H J K M linked to post
office geo (reg PCCF)

DMT Urban services

- **A Ordinary delivery--by letter carrier**
- **B Apartment buildings--served by LC**
- **E Business buildings served by LC**
- **G “Large volume receiver”--bag drop off by letter carrier**
- **M Single large PO box--bag pick up at post office**

DMT Rural services (mostly)

- **H** Rural route from urban post office
- **J** General delivery from urban PO
- **K** Group of PO boxes—in urban PO
- **T** Suburban route service—from urban PO
- **W** Rural post office—all service types
- **Z** Retired postal codes (mostly rural)

SOURCE (diagnostic)

- **F** Exact match to PCCF unique record
- **D** Match to one of PCCF duplicate records
- **C** Probabilistic match to WCF record
- **I** Pop-weighted imputation within FSA
- **2** Partial imputation from first 2 chars
- **1** Partial imputation from first character
- **0** No match even of first character

Other possible SOURCE (not generated by PCCF+)

- **G** Global positioning system (GPS)
- **P** Telephone area code + prefix
- **Q** Telephone 911
- **R** Road intersection
- **S** Street address (eg via street index)
- **T** Township, range, meridian, section (TRMS)

Explanation of the supplementary geographic codes

- RESFLG, INSTFLG, BLDGNAM*
- CPCCODE, CPCOMM*
- CSIZE, SACTYPE, NSREL, BLKURB
- HR, SUB
- QAIPPE
- EA96UID
- ER, AR, CCS, FED1996, FED2003, DPL

RESFLG (where DMT=E,G, M)

- @ Possible residence (may depend on age)
- - Improbable residence
- ? E G M but res status undertermined
- **blank** DMT not E, G or M

INSTFLG

Institutional Flag

- **E** School or university residences
- **H** Hospitals
- **I** Hospitals (only from building name)
- **N** Nursing homes
- **S** Seniors residences
- **P** Prisons, jails
- **U** Other
- **b** Not applicable (area not predominately institutional)

CPCCODE, CPCOMM

- **CPCCODE** Numeric code for Canada Post community name (sequential within province). Specific to each version of PCCF+.
- **CPCOMM** Canada Post community name (see CPCOM.CAN file). Only a sloppy, partial overlap with legal municipal (CSD) names.

Rurality-related codes

- **CSIZE** Community size
- **SACTYPE** Incl MIZ for non-CMACA
- **NSREL** North-South relationship
- **AR** Census agricultural region
- **CCS** Census consolidated subdivision
- **BLKURB*** Urban block indicator (0,1)
- **DPL*** Designated place (some prov)
- ** not well coded via postal codes*

Urban-rural continuum (CSIZE+MIZ)

Population size group of CMACA, or MIZ:

- Large metropolitan (1.25 million +)
- Medium metropolitan (500-1.24 K)
- Smaller metropolitan (100-499 K)
- Census agglomerations (10-99 K)
- Rural and small town (residual)
 - **MIZ: Strong, moderate, weak, none**

Note: These categories are also suitable for assignment from place-name based SGC coding (which can be rough--but OK for such groupings)

NSREL – North-South Relationship

- North
 - North transition
 - South transition
 - South
-
- *CSD-level variable based on 1996 census data analysis by McNiven & Puderer (2000)*

Other geographic variables

- **REG** Health region (note vintage)
- **SUB** Health district (note vintage)
- **ER** Economic region
- **EA96uid** 1996 census EA (for data)
- **FED1996** Federal electoral district
- **FED2003** (“rep order” vintage)

Little-known tidbits about PCCF+ geographic “name” files

- **CDNAMES.CAN** includes unofficial descriptive names for unnamed (numbered) census divisions (so NF Division No 10=>Labrador).
- **CDNAMES** removes redundant parts of census division names (so NB Madawaska County=>Madawaska, CDTYPE=CTY).
- **ARNAMES.CAN** includes unofficial descriptive names for unnamed census agricultural regions or crop districts (so SK Region 5A=>Yorkton).
- **HRNAMES.CAN** includes unofficial descriptive names for unnamed health regions (so NS Zone 5=Cape Breton).
- **HRNAMES.CAN** separates health region type from health region name, which considerably shortens and removes redundancy from lists of health regions (so ON Champlain District Health Council=>Champlain, HRTYP=DHC)
- **SUBNAMES.CAN** does the same thing for health district names.
- **TPHANAMES.CAN** does the same thing for Toronto public health areas.

Labelled printouts from HLTHOUT and GEOPROB files

- **MSWORD.FMT4EGEO.DOC**
Open above file, insert your HLTHOUT file, then print (without saving).
- **MSWORD.FMT4EPRB.DOC**
Open above file, insert your GEOPROB file, then print (without saving).

QAIPPE

- **Dissemination area (DA)-level**
- **Population quintiles**
- **Area-based (within CMACA or PR rural and small-town residual area)**
- **Adjusted for family size (using single-person equivalents from LICO)**
- **Also available for other census years**

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).**

Translating across vintages of census geography

- Get supplementary “translation” files:
 - EA96291, EA96286, EA96281
- Sort both files (HLTHOUT + translation) by EA96uid
- Merge by EA96uid
- Append older vintage EA__uid to file

Translating across vintages of census geography

- **DA01uid** **2001 census geography**
- **EA96uid** **1996 census geography**
- **EA91uid*** **1991 census geography**
- **EA86uid*** **1986 census geography**
- **EA81uid*** **1981 census geography**
- *** based on nearest centroid**

DIST4D.SAS

- **/* CALCULATE DISTANCES FROM EACH OF MANY EVENTS (E) */**
- **/* TO THE NEAREST SERVICES (H) BY SPECIALTY */**
- **/* READS IN A FILE OF EVENTS CODED BY PCCF+ (GEORES4D) */**
- **/* AND A FILE OF SERVICES CODED BY PCCF+ (GEOINS4D) */**
- **/* OUTPUTS A FILE OF EVENTS WITH APPENDED DISTANCES */**
- **/* TO THE NEAREST SERVICES BY SPECIALTY */**
- **/* NOTE: */**
- **/* EVENTS FILE ASSUMED TO BE OUTPUT OF GEORES4D */**
- **/* WITH SPECIALTY CODE SOMEWHERE IN FILE */**
-
- **Distance to nearest hospital with obstetrician, variable for study of birth outcomes in BC (Luo, Kierans et al, *Epidemiology* 2004);**
- **Distance to school and university participation (Frenette, *ASB* 2002)**
- **Distance to nearest hospital, distance to nearest MD (Ng et al, Amankwah)**

Evaluating coding accuracy

- **Comparisons** **Gold standard**
- **V3A vs SLI vs FSA** **1996 census EAs**
- **V4C urban vs rural** **2001 census DAs**
- **Movement of pcodes** **Distance + geog codes**

Comparison of geographic coding by 3 different methods

- **R3A** PCCF+ Version 3A
(population-weighted)
- **SLI** Single Link Indicator
(best single link)
- **FSA** Forward Sortation Area
(population-weighted)

Method of comparison

- Simple 1 % random sample of 1996 census population
- Postal codes as collected or imputed by census
- Blind coding from postal codes only, by three different methods (R3A, SLI, FSA)
- Tabulation of population in each geog unit, by the three methods
- $\% \text{ error} = \text{sum of deviations (absolute values)} / \text{pop}$ in original sample at each level
- Gold standard: EA & pop determined by census

Results of the comparison: % error at population level

<i>Geog</i>	<i>R3A</i>	<i>SLI</i>	<i>FSA</i>
• PR	0.1	0.1	0.0
• CD	0.3	0.6	0.5
• CSD	3.2	9.4	4.7
• CMA	0.2	0.4	0.3
• CT	1.9	2.7	11.6
• EA	15.8*	33.6	41.8
• DPL	20.0	50.9	30.3

• See User's Guide, Table 2 (page 17) * Larger samples better

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.

Movement of postal codes

- **Many technical changes to address ranges**
 - Usually no change of blockface or block LL
 - Very little change at higher levels (DA, CT etc)
 - Movement always within same FSA service area
- **Some reuse of retired postal codes within same FSA; if so, DMT may also change**
- **However, two complete FSAs in BC moved by Canada Post during mid-1990s**
- **Moral: Code data as received; retain results**

Observations, recommendations

- **EAs on survey files not well documented**
 - **Vintage of EA often *not* that of the survey year**
 - **EA86uid, EA91uid, EA96uid, DA01uid**
 - **Vintage of CSD, CT, HR, CSIZE etc. (even PR)**
- **Naming conventions: geoYYuid**
 - **uid => higher levels always needed with geo**
 - **YY => vintage of census geography required**

Coding Strategies

- **Addresses available**
 - **Generate postal codes from addresses**
 - **Compare to captured postal codes**
 - **Use PCCF+ to code each, and compare results**
- **Independently-coded SGC (CSD) available**
 - **CMA/CA and CD should agree**
 - **CSD should be same or adjacent**
- **Only postal codes available**
 - **Use PCCF+ diagnostics & problem output**
 - **Examine geographic codes assigned**

Thank you