Management and Aggregation of Disparate Data from Disparate Sources: Illustrations from an Evaluation of the Swiss National Science Foundation

Evaluation of STI policies, instruments and organizations:
New horizons and new challenges

Vienna, Austria

15 November 2013

Policy Motivated Reporting – Evaluation – Research

Policy Motivated Reporting

Process Monitoring

- Focus on procedures
  - Information/process bottlenecks
  - Checklist oriented

Accountability

- Typically count-based
- Key Performance Indicator (KPI) focused

Evaluation

Evaluation Functions

- Formative
- Summative
- Comparative

Evaluation Approaches

- Consumer-oriented
- Program-oriented
- Decision-oriented
- Participant-oriented
- Cultural competence & capacity building

Research

Research Functions

- Basic/fundamental
- Applied
- Practitioner

Research Designs

- Experimental
- Quasi-experimental
- Correlational
- Pre-experimental

----- RQ / Use -----

- Causal / theory testing
- Causal-comparative
- Predictive
- Descriptive

Data Organization: Reporting – Evaluation – Research

- Unit of reporting vs. unit of {datum} representation vs. unit of analysis
  - These 3 units can co-occur because they serve differing purposes/functions
- Need for rolling-up the data?
  - Move from one type of “unit” to another
  - Unit of data < unit of analysis
    - apply weighting?
  - Duplication
  - Repeated measures
  - Missing data

Project Size ≠ Data Management Complexity ≠ Analytical Complexity

Cross-sectional

- SNSF (2005-2011) rolled-up into a 2012 end point analysis (data existed for policy-based reporting & institutional study)

Longitudinal

- Longitudinal Study (1998-2007) of career and technical education in Illinois, USA (data originally existed for policy-based reporting)

Evaluations Often Balance Multiple Data Sources

- Institutional DB
  - Organizational text-based
  - Relational DB
  - Multiple flat files
- Project driven data collection
  - Survey-based
    - Custom designed but with generally unknown psychometrics
    - Off-the-shelf with known psychometrics
  - Interview
    - Single level informants
    - Multilevel informants

Data Collection & Sampling Considerations

- Known and unknown target population(s)
- Census possibility
- Probability
- Non-probability

Data Elements

Qualitative

- Orientation
  - Ethnology
  - (Reality) Correspondence
  - Social constructivism
  - Phenomenology
  - Narrative analysis
  - Ecological psychology
  - System perspective
  - Grounded Theory
- What is the intent of coding?

Quantitative

- Aggregate (composite) indices
  - Common construct
  - Distributional concerns
- Multi-level features
- Missing data

Stable Measurement is Required

Measurement is the assignment of symbols to properties of objects, or behaviors or events according to rules

- When data are aggregated from different sources, errors in measurement accumulate.
  - Measurement error does not cancel out
  - Data accumulation only makes sense when there is distributional agreement among the indicators

Examples

- Structuring SNSF Data for Analytics
- Common Merge & Unique Identifier: ISBE
- Transitioning Data from Cross-sectional to longitudinal: ISBE

Thank You

Example From SNSF Transparency Evaluation

Data Context of the SNSF Transparency Evaluation

- Unit of analysis varied
  - Many of the evaluation questions were addressed by multiple units of analysis
- Extent data analysis is the focus presented in this example

Unit of Data – Unit of Analysis: A Multiply Repeated Data Structure

- SNSF provided the Evaluation Team with an Excel Workbook representing a data extract from their institutional “Applicant” DB
- The Workbook was a variable-record file corresponding to SNSF applications (funded and unfunded).
- The primary identifier is SNFS_ID represents an APPLICATION (not an applicant)
- There were multiple records per SNSF_ID due to multiple funding years for the application with multiple disbursements possible within one year (one record)
- There were multiple SNSF_IDs per principal applicant ID due to multiple application submissions
- Some of the applications were funded, some were not within a applicant ID

Application Record Processing

- Delete records that should not be in the file
  - 37,033 rows read in from SNSF extract
  - Approximately 9,000 rows deleted per SNSF direction
  - 28,092 rows written a SAS dataset
- Begin the process of rolling-up the records
  - Horizontal roll-up
  - Vertical roll-up
- Scrubbing and recoding variables/fields prior to statistical analysis

Identify Multiple Records Per Application

- Count records by SNSF_ID and append result to record

ID_freq	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	24947	88.80	24947	88.80
2	2403	8.55	27350	97.36
3	524	1.87	27874	99.22
4	143	0.51	28017	99.73
5	47	0.17	28064	99.90
6	19	0.07	28083	99.97
7	4	0.01	28087	99.98
8	3	0.01	28090	99.99
9	1	0.00	28091	100.00
11	1	0.00	28092	100.00

Find Year of First Award

- Field array:
  - AMTGranted2004-AMTGranted2017
- Scan field array to find instance of first nonmissing data
  - Append result (FirstYR) to record
  - 28,092 rows written to a SAS dataset

Calculate Cumulative Total Award ($)

- Sum all award disbursements over all years within a record and accumulate over all records by SNFS_ID
  - Move result CUMTOTAL to the first record in an SNSF_ID set
  - 28,092 rows written to SAS dataset

Construct New Record Counter

- Generate a new random ID from a uniform distribution
- Create a new variable FUNDED
  - if CUMTOTAL>0 then FUNDED=1 else FUNDED=-1
  - 28,092 rows written to SAS dataset

Trim Records for Inclusion (time) Period

- Trim all records (applications) with a start date before 2005
  - 1353 Cases rejected
  - 26,739 rows written to SAS dataset

Split APPLICANT Dataset from APPLICATION Dataset

- Selected from APPLICATION dataset the first record of a case and where APPLICANTGENDER > “”
- Save result into APPLICANT dataset
  - 26,739 record written to SAS dataset

Construct/Scrub Various Applicant-based Variables

- AGE: Based on DOB, filters (25<= AGE <90) retains in the data
- Dummy code GENDER: Male=1, Female=0
- Submission DIVISION: ADMINDIVISION was used to create a new variable DIVISION with 5 levels (was over parameterized {5 dummy variables} so that any subset could be examined depending on which dummy code was omitted.
  - Division 1: D1=1 D2=0 D3=0 D4=0 D5=0
  - Division 2: D1=0 D2=1 D3=0 D4=0 D5=0
  - Division 3: D1=0 D2=0 D3=1 D4=0 D5=0
  - Careers: D1=0 D2=0 D3=0 D4=1 D5=0
  - Other: D1=0 D2=0 D3=0 D4=0 D5=1

Continued

- Invoke stronger time inclusion filter: only 2005-2012
- set up dummy code with referent as “Kantonale Universitat”
  - Kantonale: INST1=0 INST2=0 INST3=0
  - ETH: INST1=1 INST2=0 INST3=0
  - Fach INST1=0 INST2=1 INST3=0
  - Andere INST1=0 INST2=0 INST3=1
- NEWAPPLICANT indicator created to differentiate new applicants (PI’s) from previous SNSF grant applicants (regardless of prior funding success)
  - 26,418 record/cases written to SAS dataset

And then the statistical analyses began

return

Problems with (supposedly) Unique Identifier Crossing Different Data Sources

Longitudinal Study (1998-2007) of career and technical education in Illinois, USA (data exist for policy-based reporting)

Wage Data Issues

- The only identifier on the UI wage data is SSN.
- Frequency table indicates that there are SSNs associated with multiple employers in one quarter.

Year by Quarter

Year	Quarter				Total
	1	2	3	4	
1999	84	249	201	174	708
2000	139	159	134	116	548
2001	90	97	93	81	361
2002	70	76	74	68	288
2003	53	65	62	56	236
2004	49	55	57	53	214
2005	45	49	41	44	179
Total	530	750	662	592	2534

For a particular SSN xxx-xx-7869

In 2nd quarter, 1999, this SSN got paid

by 249 different employers.

Looking at those 249 wage records

year	quarter	ssn	UIACCT	FEIN	wages	NAICS
1999	2	xxx-xx-7869	4179639	0	263	
1999	2	xxx-xx-7869	548440	132677117	1179	722110
1999	2	xxx-xx-7869	2035483	133581452	832	561720
1999	2	xxx-xx-7869	2045828	222228945	3641	
1999	2	xxx-xx-7869	1233748	222623485	788	326112
1999	2	xxx-xx-7869	2047476	223144609	100	722211
1999	2	xxx-xx-7869	4215990	223606735	41	561320
1999	2	xxx-xx-7869	4206352	232949247	182	561730
1999	2	xxx-xx-7869	735890	310986349	1994	722211
1999	2	xxx-xx-7869	1177019	351604308	517	722211
1999	2	xxx-xx-7869	4109149	351967143	2285	722110
1999	2	xxx-xx-7869	2706	360896040	3075	813410
1999	2	xxx-xx-7869	10378	361140620	2980	332116
1999	2	xxx-xx-7869	185268	361349898	5423	333514
1999	2	xxx-xx-7869	16986	361546460	474	311812
1999	2	xxx-xx-7869	15749	361562080	2910	713910
1999	2	xxx-xx-7869	13746	361785860	3909	713910

Sum the wages, this SSN earned $497,443 in 2nd quarter, 1999.

The SSN in the BASE data set

- Look up this SSN xxx-xx-7869 in **ISBE** data
- 2 records were found, these 2 records have different first and last names, but same SSN.

subj	lname	fname	ssn
150162990000425000000032051901	AA	AA	xxx-xx-7869
150162990000412500000036149930	BB	BB	xxx-xx-7869

address	city	grade96	grade97	grade98	grade99	grade00	grade01
8104 S HOUSTON AV	CHICAGO	.	.	.	.	10	11
2743 N AUSTIN AV FL	CHICAGO	.	.	.	.	.	.

grade02	grade03	grade04	grade05	grade06	combine
.	.	.	.	.	AAAAxxx-xx-7869
.	.	.	10	.	BBBBxxx-xx-7869

Scrub SSN in ISBE Student Data

- Purpose: need to merge Wage data with ISBE student data
- Since SSN is the only identifier on Wage data, we need to merge the two datasets by SSN
- In ISBE data not all the students have SSN, some have blanks, some have 000000000.

return

Transitioning Data from Cross-sectional to Longitudinal

Longitudinal Study (1998-2007) of career and technical education in Illinois, USA (data exist for policy-based reporting)

Building the Longitudinal Panel

Considerations

1) Every new year we have a panel expansion
2) We are not fully capturing three different subclasses of students: “Browsers” and “Explorers” and “Movers”

   Explorers—1 year in CTE

   Browser—2 years in CTE

   Concentrator—3 or 4 years in CTE

Solution: Create a longitudinal panel that includes all students from 1996-2006

Issues

New student ID or possible data entry error?

(similar subj, the red part indicates these 2 records are from the same school)

subj	lname	fname	ssn	address	City	Grade 97	Grade 98	Grade 99
140162000000113000000000041001	lname1	fname1	xxx-xx-5771	675 W LAKE ST	OAK PARK			12
140162000000113000000004001001	lname1	fname1	xxx-xx-5771	675 W LAKE ST	OAK PARK	10	11	

Movers

(Different schools—as indicated by school id, but same last first names, same SSN)

subj	lname	fname	ssn	address	City	Grade 00	Grade 01	Grade 02
150162990001725000000032323219	lname2	fname2	xxx-xx-7022	5919 S JUSTINE ST	CHICAGO		10	11
140162310000116000000200300189	lname2	fname2	xxx-xx-7022	3158 WEST 88TH ST APT 201	EVER GREEN PARK	9		

Enrolled at two different schools

(Same school district, one High school records, one Voc Center records)

subj	lname	fname	ssn	address	City	Grade 98	Grade 99	Grade 00	Grade 01
170640870000125000000002044473	lname3	fname3	xxx-xx-3048	706 S ALLIN	BLOOMINGTON	9	10		
170640870410141000000000001353	lname3	fname3	xxx-xx-3048	705 S ALLIN	BLOOMINGTON			10	11

Build New 'Key' Variable

- Purpose: To find/create a variable that can uniquely identify a student in the dataset
  - Possible variable: SSN but there are missing values, 000-00-0000, recycled, sharing
  - Last name, first name, city, address, street number
- New Key /Identification Variable:
  - 1) Combine1= last name||first name||ssn
    - (if ssn≠ blank or ssn ≠ 000000000)
  - 2) Combine2= last name||first name||city||street number
    - (if ssn= blank or ssn = 000000000)

Process

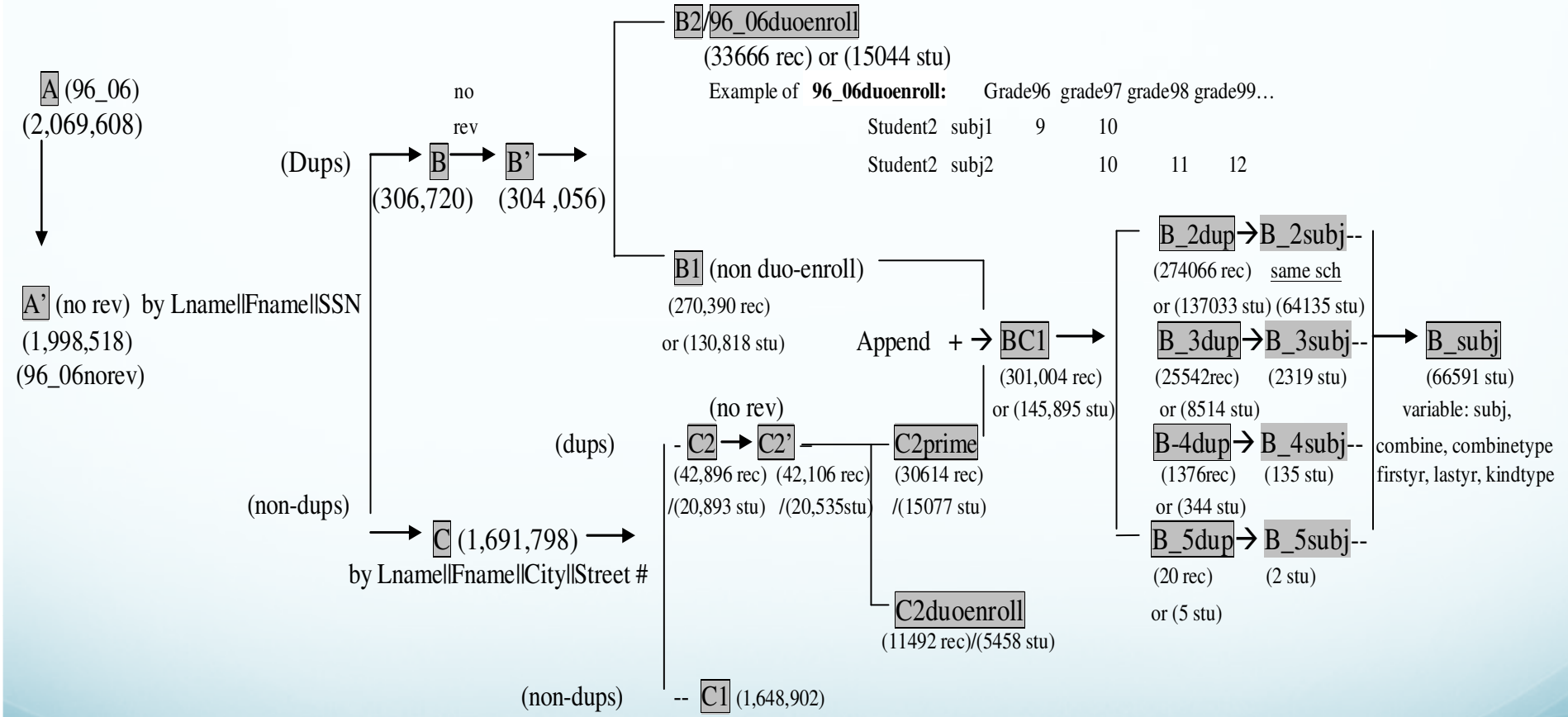
1. Merge datasets from 96, 97, 98, 99, 00, 01, 02, 03, 04, 05, 06 by subj
2. Eliminate reverse graders

Example of reverse graders

lname	fname	ssn	Grade 96	Grade 97	Grade 98	Grade 99	Grade 00
lname4	fname4	xxx-xx-9921		10	11	10	12
lname5	fname5	xxx-xx-0936	9	10	9	10	

3. Identify duplicated record by combine1 AND combine2

Processing flow chart



Examples

Dual-enrolled (dataset-B2 in flow chart, temporarily set aside)

subj	lname	fname	ssn	address	City	Grade 02	Grade 03	Grade 04	Grade 05
03025040000012600000011250000	lname5	fname5	xxx-xx-8394	1202 N MERCHANT #8	EFFINGHAM	9	10	11	12
11015801060096000039357788394	lname5	fname5	xxx-xx-8394	1202 MERCHANT ST #8	EFFINGHAM		10		12

Duplicated, non dual-enrolled, different school ID (temporarily set aside)

subj	lname	fname	ssn	address	City	Grade 00	Grade 01	Grade 02
19022000040014000000100220025	lname6	fname3	xxx-xx-2082	213 S CLAY	HINSDALE		11	12
19022086000011700000000220025	lname6	fname3	xxx-xx-2082	2505 W. 35TH ST.	OAK BROOK	10		
14016214000071700000705011007	lname7	fname7	xxx-xx-0781	3700 MEADOW DR	ROLLING MEADOWS	10 (03)		12 (05)
14016512046014600000705011007	lname7	fname7	xxx-xx-0781	3700 MEADOW DR	ROLLING MEADOWS		11 (04)	

Duplicated, non dual-enrolled, same school ID (roll-up)

subj	lname	fname	ssn	address	City	Grade 99	Grade 00	Grade 01
14016203000011700000007020001	lname8	fname8	xxx-xx-3454	635 KNOX AVENUE	WILMETTE	9		
14016203000011700000020020001	lname8	fname8	xxx-xx-3454	635 KNOX AVE	WILMETTE		10	11

Examples

Records on B_2subj, B_3subj, B_4subj, B_5subj

Dataset	Combine	Grade96	Grade97	Grade98	Grade99	Grade00	Grade01	Grade02	Grade03	Dup #
B_2subj	lname9fname9xxx-xx-0028	.	.	9	.	.	.	.	.	2
	lname9fname9xxx-xx-0028	.	.	.	10	.	12	.	.	
B_3subj	lname10fname10xxx-xx-0666	9	10	.	.					3
	lname10fname10xxx-xx-0666	.	.	.	12					
	lname10fname10xxx-xx-0666	.	.	11	.					
B_4subj	lname11fname11xxx-xx-4284	.	.	.	.	.	10	.	.	4
	lname11fname11xxx-xx-4284	.	.	.	.	9	.	.	.	
	lname11fname11xxx-xx-4284	.	.	.	.	.	.	.	12	
	lname11fname11xxx-xx-4284	.	.	.	.	.	.	11	.	
B_5subj	lname16fnam16exxx-xx-0758	.	.	.	.	.	12	.		5
	lname16fname16xxx-xx-0758	.	.	.	.	11	.	.		
	lname16fname16xxx-xx-0758	.	.	.	.	.	.	12		
	lname16fname16xxx-xx-0758	.	.	.	10	.	.	.		
	lname16fname16xxx-xx-0758	.	.	9	.	.	.	.		

Examples

Records on B_subj (these are the records were rolled-up based on same combine and same school ID in multiple records)

kindtype	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00	63000	94.61	63000	94.61
000	2313	3.47	65313	98.08
0000	135	0.20	65448	98.28
00000	2	0.00	65450	98.29
33	48	0.07	65498	98.36
44	1033	1.55	66531	99.91
444	4	0.01	66535	99.92
66	54	0.08	66589	100.00
666	2	0.00	66591	100.00

33—duplicate 2 times, both records are from special schools

444—duplicate 3 times, all records are from voc centers

666—duplicate 3 times, all records are from special ed coop schools

Building The Longitudinal Dataset

N = 1,715,493

(N_{nondup}= 1,648,902; $N_{roll\text{-}up}$= 66,591)

Index variables:

Yrcount—number of years appeared in dataset

Firstyr—first year of appearance in dataset

Lastyr—last year of appearance in dataset

Bgrade—beginning grade when first appeared in dataset

Egrade—ending grade when last appeared in dataset

Repeat– (0/1) whether repeat a grade or not (eg. …99… or …9..9…)

Grade Patterns on New Dataset

(Cases with 4 years of records & increasing grade)

Pattern	Actual Grade Pattern	Year	N of New	N of Old	Difference (Old – New)
9-10-11-12	…….9101112	03-06	17,524	18,947	1,423
	……9101112.	02-05	17,515	18,960	1,445
	…..9101112..	01-04	18,025	19,739	1,714
	….9101112…	00-03	17,765	19,471	1,706
	…9101112….	99-02	16,255	17,768	1,513
	..9101112…..	98-01	12,821	13,573	752
	.9101112……	97-00	12,985	13,265	280
	9101112…….	96-99	13,890	15,206	1,316

Grade Patterns in New Dataset
(Cases with 3 years of records)

Year	Pattern	N (new)	Pattern	N (new)
03-06	…….9.1112	7513	…….910.12	5,436
02-05	……9.1112.	7409	……910.12.	4,976
01-04	…..9.1112..	7699	…..910.12..	5,230
00-03	….9.1112…	7629	….910.12…	5,467
99-02	…9.1112….	4634	…910.12….	5,191
98-01	..9.1112…..	4840	..910.12…..	4,004
97-00	.9.1112……	7232	.910.12……	4,467
96-99	9.1112…….	5821	910.12…….	5,507

These students have enrolled 3 of 4 years in CTE, and they have graduated within 4 years without repeating a grade or skipping a grade.

Grade Patterns on New Dataset

(Cases with 3 years of records)

Year	Pattern	N (new)	Pattern	N (new)
03-06	…….91011.	5670	…….910.12	6,753
02-05	……91011..	6053	……910.12.	6,866
01-04	…..91011...	5837	…..910.12..	7,082
00-03	….91011….	5468	….910.12…	7,983
99-02	…91011…..	5080	…910.12….	9,443
98-01	..91011…...	4644	..910.12…..	7,669
97-00	.91011…….	3741	.910.12……	5,343
96-99	91011……..	6115	910.12…….	6,149

Exclude

Repeaters: ..991011.. or …9910.12…

Skippers: …91112…. or ….911.12…

Grade Patterns in New Dataset

(3 years of records-comparison)

Year	9-.-11-12	9-10-.-12	9-10-11-.	.-10-11-12	N of new 3/4	N of old 3/4	Difference
03-06	7,513	5,436	5,670	6,753	44,319	49,699	5,380
02-05	7,409	4,976	6,053	6,866	44,264	50,789	6,525
01-04	7,699	5,230	5,837	7,082	45,587	52,474	6,887
00-03	7,629	5,467	5,468	7,983	46,018	53,599	7,581
99-02	4,634	5,191	5,080	9,443	42,116	50,231	8,115
98-01	4,840	4,004	4,644	7,669	34,730	39,550	4,820
97-00	7,232	4,467	3,741	5,343	34,048	41,104	7,056
96-99	5,821	5,507	6,115	6,149	38,798	46,402	7,604

These students have enrolled 3 of 4 years and they have graduated within 4 years without repeating a grade or skipping a grade

Grade Patterns on New Dataset

(2 out of 4 years of records—new capture)

Year	9-10-.-.	9-.-11-.	9-.-.-12	.-10-11-.	.-10-.-12	.-.-11-12	N new 2/4
03-06	9387	4016	6482	3322	4397	8539	36143
02-05	9124	3926	6074	3654	3959	8139	34876
01-04	8836	4016	6766	3499	4092	8396	35605
00-03	8702	3908	6360	3568	4057	7522	34117
99-02	9239	2481	4155	4428	4574	8807	33684
98-01	7794	2682	4371	3966	3828	11017	33658
97-00	9557	3081	6634	2502	3029	8885	33688
96-99	11003	3254	5715	4094	3896	7756	35718

Cohort on New Dataset

(Year 96-99)

96	97	98	99	Pattern	N of new	N of Old ??
9	10	11	12	4 year	13,890	15,206
9	.	11	12	3 out of 4 year	5,821	46,402
9	10	.	12		5,507	
9	10	11	.		6,115	
.	10	11	12		6,149	
9	10	.	.	2 out of 4 year	11,003	Not captured
9	.	11	.		3,254	
9	.	.	12		5,715	
.	10	11	.		4,094	
.	10	.	12		3,896	
.	.	11	12		7,756	
9	.	.	.	1 out of 4 year	24,135	Not captured
.	10	.	.		11,001	
.	.	11	.		7,256	
.	.	.	12		15,605	

Remaining Questions

how to categorize patterns like:

Repeaters

- …99……, ..1212…….
- Repeat more than 2 years: …999….. , …91011121212..
- Repeat more than 2 grades: ..9910111112…

Skippers

- …991112…. , …9101012…..
- Possible skippers: ….9.12……

Other Patterns

- …9.10…… , …9…101112.

return