The Soft Neighborhood Model: Guide to the PPS Data Set

Brooke Cowan
Matt Marjanović
brooke@softneighborhoodmodel.org
maddog@softneighborhoodmodel.org
http://www.softneighborhoodmodel.org/

November 15, 2015

1 Introduction to the Data Set

In *The Soft Neighborhood Model: A Dynamic Enrollment-Balancing Framework*, we apply the Soft Neighborhood model to real PPS data provided to us by the district. We have made this data set publicly available and are providing this guide for anyone who wishes to explore it. The data set contains seven years of PPS student data (2008/09–2014/15) in which addresses have been anonymized. While there are clear limitations and flaws in this data set, to the best of our knowledge there aren't any others that are similarly comprehensive and free of encumbrances.

We hope that by releasing this data, we can contribute to the empowerment of PPS stakeholders and the transparency of PPS initiatives. The 2015–16 District-Wide Boundary Review is one such initiative. We developed the Soft Neighborhood model as an alternative to the classic hard boundary model that is current practice in PPS, and we used this PPS-provided data set to test our model. With the release of this data, anyone who has a solution to enrollment balancing can implement and test their ideas. Access to a common data set empowers all of us and testing against a common data set is the only valid scientific way to compare competing solutions to the same problem. The two scenarios published by PPS on October 29, 2015 represent one such competing solution. The PPS data set may be useful to those folks who are trying to independently verify the PPS-produced statistics and analysis accompanying these scenarios.

The data provided by PPS is composed of two sub-data sets, the STUDENTS data set and the SCHOOLS data set. The STUDENTS data set contains historical and current information from the last seven school years. The SCHOOLS data set contains information about K–8, K–5, and 6–8 schools in PPS in the year 2014-15. In Section 2, we give a detailed overview of what's in these two data sets. In Section 3, we explain what's *not* in them. Lastly, in Section 4, we suggest corrections that would make the data sets even more useful to the community. We encourage anyone requesting corrections from PPS staff to use our suggestions as talking points in their communications with the district.

2 What's in the Data Set

2.1 The Students Data Set

The STUDENTS data set consists of 208,394 records for students in kindergarten through eighth grade over the seven-year period from 2008–2015. Table 1 shows the number of student records (K–8 and K only) each year. Only students attending neighborhood schools (but not necessarily neighborhood programs!) are included in this data set — see Section 3 for more details on the limitations of this data set.

Table 1: The number of student records (K–8 and K only) per year in the Students database.

Year	**K–8**	**K**
2008-09	28,760	3,601
2009-10	29,077	3,726
2010-11	29,211	3,657
2011-12	29,875	3,715
2012-13	30,056	3,884
2013-14	30,671	3,843
2014-15	30,744	3,681
2008-15	208,394	26,107

The Students data is stored in CSV format in the file named `DBRAC Data request on student points and school points v3 STUDENTS.csv`. The first line of text contains the field names; the remaining lines are student data, one line per (student, year) record. The contents of these records are detailed in Table 2.

Here's a sample row from the data set:

```
SCH_YEAR,Grade,campus_enrolled,s_e_campus_code,ES_NAME11,capture,X_COORD,Y_COORD
...
2010-11,5,Atkinson,828,Harrison Park,,7674958,677065
...
```

This row represents a student who in 2010–11 was in the 5th grade at Atkinson Elementary (School ID 828). This student's default neighborhood school would have been Harrison Park, and therefore the `capture` field is blank (false). The anonymized X and Y coordinates of this student's home address are (7674958, 677065) in the Oregon State Plane coordinate system.

2.2 The Schools Data Set

The Schools data set contains information about K–8, K–5, and 6–8 schools in PPS during the year 2014–15. The Schools data is stored in CSV format in the file named `DBRAC Data request on student points and school points v3 SCHOOLS.csv`. The first line of text contains the field names; the remaining lines are school data, one line per school record. The contents of these records are detailed in Table 3.

3 What's Not in the Data Set

The data set is missing some critical information, severely limiting the analysis that can be carried out using it. As we have noted in our paper, these limitations impact the extent to which we have been able to validate the Soft Neighborhood model. The limitations also impact the degree to which the community can use them to "sanity check" the district's

Table 2: Fields in the Students data set.

Field Name	**Concept**	**Description**
SCH_YEAR	School Year	The school year for this record (ranging from 2008-09 through 2014-15)
Grade	Grade	Student's grade level (K through 8) during that school year
campus_enrolled	Assigned School	Student's assigned school during that school year. Schools included in the data set are K–8, K–5, and 6–8 schools that have a neighborhood program on their campus. District-only focus options, alternative, and charter schools are not included.
s_e_campus_code	Assigned School ID	A numeric ID associated with the assigned school
ES_NAME11	Capture School	The name of the neighborhood school which captures the student by default. Values for this attribute can be the name of a PPS neighborhood school for in-district students (205,113 records)), out of district (3,264 total records) for out-of-district students, or not yet assigned (18 total records, not clear when this is used).
capture	Capture?	1 if the assigned school is the same as the default capture school (where the student lives); blank (empty string) otherwise (e.g., transfer students)
X_COORD, Y_COORD	Home Address	The location of the student's home, in Oregon State Plane coordinates, presumably in US Survey Feet. To anonymize the data, each coordinate in each record has been independently offset with a random number. (We think the random offsets are drawn from a range of plus/minus a couple hundred feet, but we're not sure.)

Table 3: Fields in the SCHOOLS data set.

Field Name	**Concept**	**Description**
GRADE	Grade Configuration	K-8, K-5, or 6-8
OBJECTID	*unknown*	Probably some kind of surrogate key.
NAME	School Name	The name of the school or campus if the school comprises more than one building.
ADDRESS	Street Address	The address of the school/campus
CITY	City	The city of the school/campus (Portland)
STATE	State	The state of the school/campus (OR)
ZIPCODE	Zip-Code	The zip-code of the school/campus
LEVEL_NO	Grade Configuration ID	2 for a 6–8 grade configuration, and 1 otherwise
LEVEL_	Grade Configuration	Middle or Junior High for a 6–8 grade configuration, and Elementary otherwise
POINT_X, POINT_Y	School Address	The coordinates of the school's address represented in Oregon State Plane coordinates, North Zone, in feet (probably US survey feet)
pps_short_name	School Name Alt	Shortened name of the school
style2012_13	*Unknown*	*Unknown*
program_id	School ID	Numeric ID representing the school or campus. This numeric ID corresponds more or less to the se_campus_code field in the STUDENTS data set.
P_to_K	Grade Configuration?	Repeat of GRADE?
KG_HR_cnt	Kindergarten Sections	Number of kindergarten homeroom classrooms for 2014-15 (nominally; not sure what happens with blended sections, etc.)
enroll	Total Enrollment	Total enrollment at the school
DC_enroll	Direct Certification Enrollment	Number of low SES students by direct certification at the school
Asian_enroll	Asian Enrollment	Number of Asian students at the school
Black_enroll	Black Enrollment	Number of Black students at the school
Hispanic_enroll	Hispanic Enrollment	Number of Hispanic students at the school
Multiple_enroll	Multiple Race Enrollment	Number of multiple race students at the school
Native_enroll	Native American Enrollment	Number of Native American students at the school
PacIsl_enroll	Pacific Islander Enrollment	Number of Pacific Islander students at the school
White_enroll	White Enrollment	Number of White students at the school

numbers and claims with respect to the District-Wide Boundary Review scenarios. However, since it may be the best resource available for this purpose, we encourage community members to use it and to fully understand its flaws. We also encourage people to request that district staff update the data set to make it more useful.

Here are the major flaws in this data set. Most of these are discussed in our main Soft Neighborhood paper as well.

1. **No correlation of student records from year to year:** There is no identifier which would allow one to figure out which student records (over multiple years) correspond to the same student.

2. **Lack of consistency in student addresses:** The coordinates for each home address have been randomly perturbed — but a different random offset has been chosen for each *record*, not each *address*. (For example, the coordinates for "4120 NE 22nd Ave." may look like "4205 NE 23rd Ave." in one record, but will be something else in another.)

3. **No sibling information**: Sibling relationships are not explicitly available in the data set, nor is it possible to guess sibling relationships by looking at addresses, due to the lack of consistency in student addresses discussed above. This makes it impossible to evaluate the effect of guaranteed placement of co-enrolled siblings at the same school in the Soft Neighborhood model, or the effect of grandfathering siblings in the PPS scenarios.

4. **Lack of information about new vs continuing students**: There is no explicit information in the data set about which students are new and which students are continuing from an earlier grade. Furthermore, lack of consistency in address randomizing (item 2) makes it impossible to guess this information. For the Soft Neighborhood model, this prevents analyzing how well the model is able to maintain balanced enrollments as students move up from kindergarten through the primary and middle grades (which is why we constrained our historic analysis to kindergarten only).

5. **Lack of explicit non-capture enrollment qualities** There is no differentiation of lottery transfers from petition transfers, nor assignment due to being co-enrolled younger sibling.

6. **Lack of sufficient information about focus options and immersion programs:** This data set contains no information about students enrolled in any of the focus-option or immersion programs which have no neighborhood preference and are not co-located with a neighborhood program; those students are simply absent from the data set. Conversely, students who attend non-neighborhood programs with a neighborhood preference (or co-location with a neighborhood program) *are* included in the data set, but are not distinguished from the students in the neighborhood programs. *FOR EXAMPLE,* there are no Winterhaven students in the data set; on the other hand, everyone from Woodstock is there (we think?), but there is no way to tell who is in Mandarin Immersion versus the non-immersion population. This makes it impossible to estimate the effect of program placement or the role that focus-option/immersion programs play in district-wide enrollment balancing.

7. **No student-specific socio-economic and racial/ethnic information**: This information was deemed too sensitive by the district from the standpoint of privacy and therefore impossible to release. We find this reasonable, but it does prevent direct evaluation of the effect of any proposed model on racial/SES mixing and diversity.

8. **Lack of real information regarding target capacities**: The data set itself does not contain information regarding target capacities at each school. The SCHOOLS data set does report the number of kindergarten classrooms at each school in 2014-15. However, this information does not generalize well to earlier years, and it does not capture whether those numbers are an artifact of overcrowding or underenrollment (e.g., district was forced to convert some space into an extra classroom). PPS has released some detailed information about building capacities that could potentially be incorporated into the dataset to improve its effectiveness.

This is not an exhaustive list of all of the issues we found with the data while working with it, but it pretty accurately covers the major issues. We welcome questions about the data from anyone trying to work with it since we very well may have encountered the same problem but failed to mention it here.

4 Improving the Data Set

The following corrections to the STUDENTS data set would allow for complete validation of assignment models without compromising student privacy and anonymity:

1. **Data for all programs, and complete program identification**: The data set should include data for the student population of all programs, e.g., focus option schools/programs as well as neighborhood schools. Furthermore, student data should identify the specific program in which a student is enrolled (e.g., *Mandarin Immersion* or *Odyssey*), not just the campus. This would allow one to tease apart the populations attending such programs, and also allow for analyzing how focus-option populations interact with the distribution of neighborhood school populations.

2. **Consistently anonymized addresses**: Ideally, when a given address (in this case, a coordinate pair of (feet-east, feet-north)) is randomly perturbed to anonymize it, the same perturbed result should be used for *every* instance of it. For instance, if the address (84757.1E, 33364.2N) is tweaked to (84790.2E, 33298.4N) the first time it is encountered, then it should be tweaked to that same value everywhere in the data set. This would allow one to make educated guesses as to which students are newly-enrolled at a school, and which students are siblings.

3. **Consistent student identifier**: Additionally, the data would be even more clear if each student were tagged with a unique-ID that was consistent from one year to the next. The actual PPS student-ID should certainly remain private, but a one-way hash of that ID could keep it private and serve the same purpose in the data set. This would make it very clear which students were continuing on from one year to the next.

4. **Explicit sibling references**: Consistent student ID's would also allow for an explicit reference to the next older sibling and/or same-year sibling (if any) — which would clarify which students should be handled as co-enrolled siblings and how to handle them.

We respectfully request that anyone communicating to district staff asking for corrections to this data set do so in a way that is consistent with our recommended fixes as outlined above. Hopefully, we can thereby obtain a data set that is maximally useful for the community. We also encourage anyone interested to use the data and report results back to us. We'd love to know what people are able to do with this data!