

PEN Data Cleaning: The Bug Report

<i>Section 1: Background</i>	<i>0</i>
<i>Section 2: Reading a bug report</i>	<i>3</i>
<i>Section 3: Addressing a bug report</i>	<i>4</i>

Summary

This document is written to help you read and address a bug report. A bug report is a listing of variables and observations that are missing, inconsistent, or could potentially be data entry errors. The document has three sections; section 1 is a background to the PEN data structure. Section 2 shows how to read a bug report and section 3 shows how to fix the bugs.

Section 1: Background

Each database contains a number of tables and to the extent that it was possible, we tried to ensure that each table corresponded to a page or section of the questionnaire.

As an example, consider village survey 1 (V1). An extract of part of the cover page is shown below

Village Survey 1 (V1)

Note: See the Technical Guidelines for the appropriate source of information and respondents for the various questions in the village surveys.

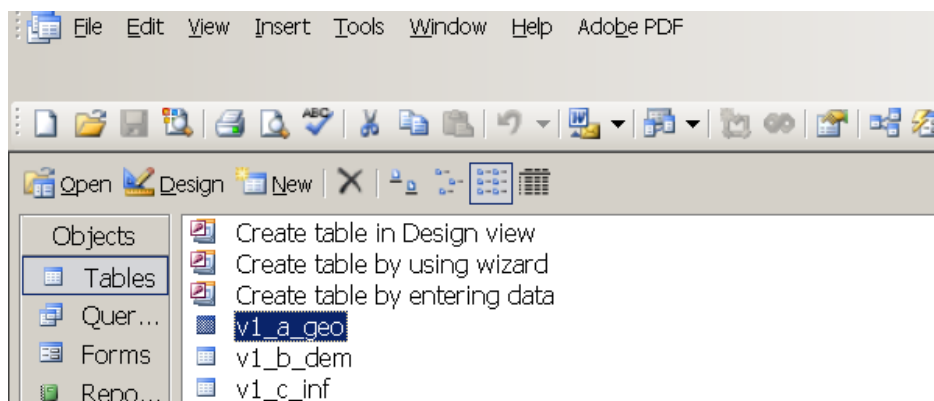
Control information

Task	Date(s)	By who?	Status OK? If not, give comments
Meeting with officials			
Village/focus group meetings			
Other interviews			
Checking questionnaire			
Coding questionnaire			
Entering data			
Checking & approving data entry			

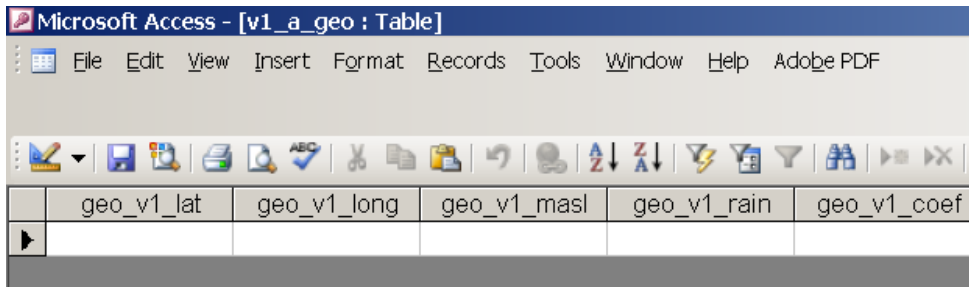
A. Geographic and climate variables

1. What is the name of the village?	1. _____ (name)	2. _____ (village code)
2. What are the GPS coordinates of the centre of the village? (UTM format)		
3. What is the latitude of the village?	degrees	
4. What is the longitude of the village?	degrees	
5. What is the altitude (masl) of the village?	masl	
6. What has been the average annual rainfall (mm/year) in the district during the past 20 years (or less, see guidelines)?	mm/year	

The corresponding database is a collection of tables which can be viewed in the tables section.



The table names have the structure [Module][Section][About] so for example, **v1_a_geo** corresponds to *Village1_SectionA_and is about Geographic and climatic variables*. This table has a number of variables (fields) which correspond to questions in the questionnaire. As an example, v1 has these variables/fields



Which correspond to Questions 3,4,5,6, and 7 of section A of this V1

A. Geographic and climate variables

1. What is the name of the village?	1. <i>(name)</i>	2. <i>(village code)</i>
2. What are the GPS coordinates of the centre of the village? (UTM format)		
3. What is the latitude of the village?		<i>degrees</i>
4. What is the longitude of the village?		<i>degrees</i>
5. What is the altitude (masl) of the village?		<i>masl</i>
6. What has been the average annual rainfall (mm/year) in the district during the past 20 years (or less, see guidelines)?		<i>mm/year</i>
7. What is the coefficient of variation in rainfall for the past 20 years? <i>(Note: To be filled in if data are readily available.)</i>		

So a bug report would be looking for missing data and inconsistencies with these variables

Section 2: Reading a bug report

For each of the problem tables, I start by producing a codebook (this will help you know what variables have problems). So for example here is a table in V1

Data Summary for v1_a_geo.dta

Variable	Obs	Unique	Mean	Min	Max	Label
...						
geo_v1_ut~zn	9	1	.	.	.	utm zone
geo_v1_u~stn	9	9	651350.8	643529	660133	utm easting
geo_v1_u~rtn	9	9	1702478	1669922	1721980	utm northing
...						
geo_v1_rain (mm/year) in	10	6	1741.3	1000	2300	average annual rainfall

Missing Data for v1_a_geo.dta

Note: Look for the Missing cases column wise

village	villcode	geo_v1_utm_zn	geo_v1_utm_estn	geo_v1_utm_nrtn	geo_v1_rain
Chuiguarabal	11	.	.	.	1000
Las Ventanas	13	15	648924	1669922	.
Quiquibaj	14	15	643529	1673579	.
El Rancho	15	.	.	.	2300
Los Ramírez	18	.	.	.	2000

This is what part of the cleaning report for v1_a_geo would look like. The first table above is a compact codebook that shows the variable names and labels (last column) so for example the variable geo_v1_rain is the average annual rainfall which corresponds to question 6 of table A in the V1.

The second table is a listing of problem cases, so you can see that geo_v1_utm_zn which is the utm zone is missing for the villages Chuiguarabal, El Rancho, and Los Ramírez. These need to be entered.

You will also notice that the villages Las Ventanas, Quiquibaj are also listed yet they do not have data missing for the UTM zone. However when you look at the entire table, you can see that Las Ventanas, Quiquibaj are missing rainfall data and hence the dots for geo_v1_rain.

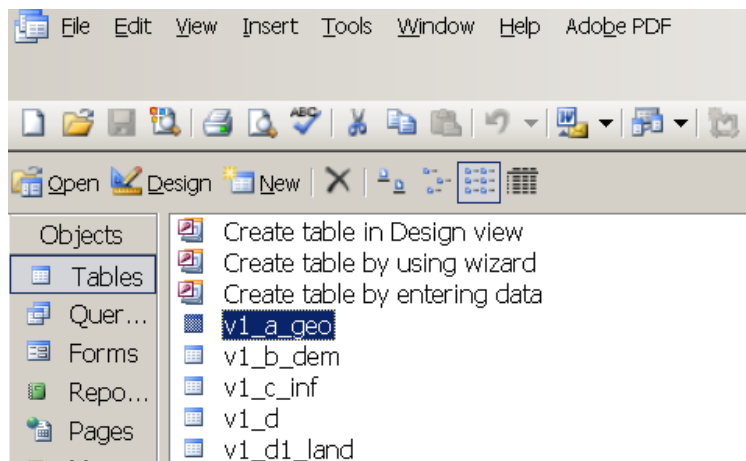
Section 3: Addressing a bug report

Given the bug report for v1_a_geo

Missing Data for v1_a_geo.dta

village	villcode	geo_v1_utm_zn	geo_v1_utm_estn	geo_v1_utm_nrtm	geo_v1_rain
Chuiguarabal	11		.	.	1000
Las Ventanas	13	15	648924	1669922	.
Quiquibaj	14	15	643529	1673579	.
El Rancho	15		.	.	2300
Los Ramírez	18		.	.	2000

1. Find the questionnaire for the village with a problem. In this case, we would be looking for the questionnaire for the village Chuiguarabal
2. Open the access database for v1



3. Open the table with errors (i.e. v1_a_geo) by double clicking on it.

The screenshot shows the Microsoft Access application window displaying the 'v1_a_geo' table. The table has six columns: offyr, offmon, offday, offby, and fgyr. The data is as follows:

	offyr	offmon	offday	offby	fgyr
	2004	11	20	9	2005
	2004	11	20	9	2005
	2004	11	20	9	2005
	2004	11	20	9	2005

4. Select the field with the village code (or household code if you are working with household data).

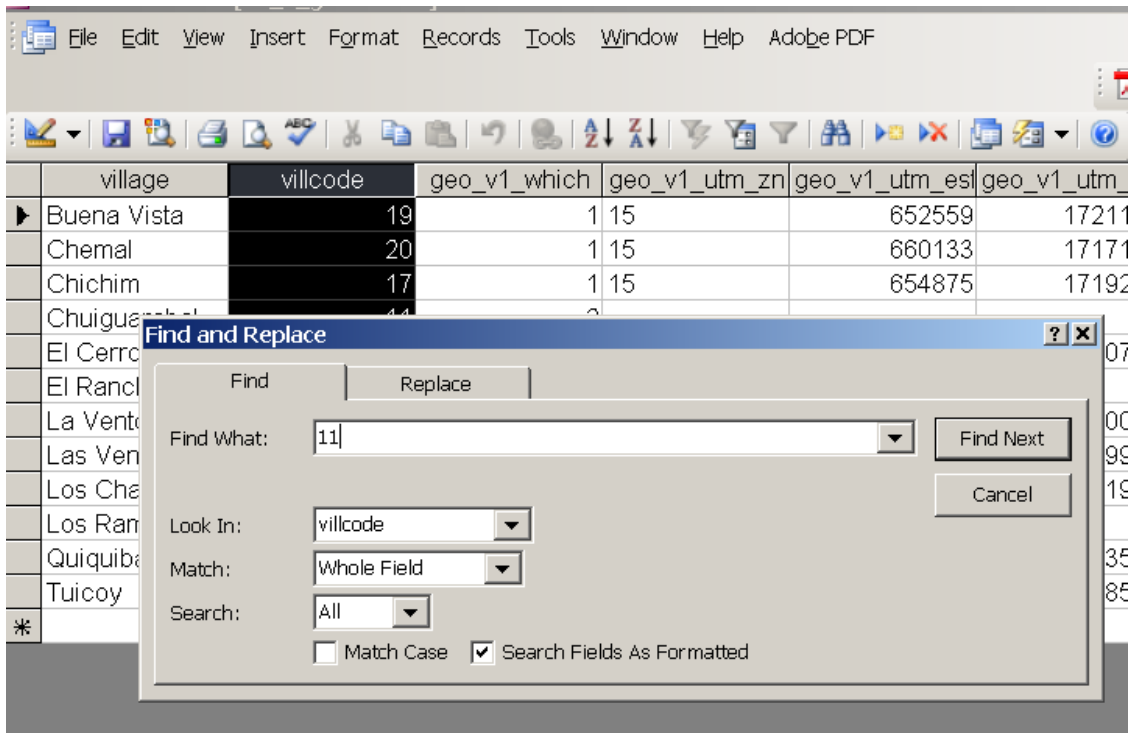
	village	villcode	geo_v1_which	geo_v1_utm_zn	geo_v1_utm_esl
▶	Buena Vista	19	1	15	652559
	Chimal	20	1	15	660133
	Chichim	17	1	15	654875
	Chuiguarabal	11	2		
	El Cerro	12	1	15	644906
	El Rancho	15	2		
	La Ventosa	16	1	15	657115
	Las Ventanas	13	1	15	648924

5. With the village code field (villcode) selected, find the problematic village. This will help you narrow in on the data for this village. As per our example we will be looking for villcode 11

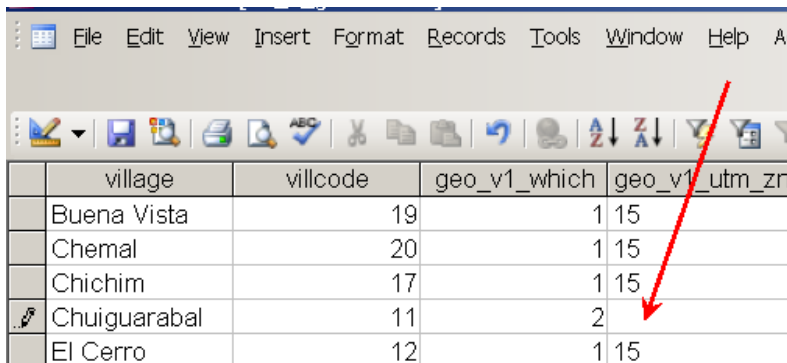
Microsoft Access - [v1_a_geo : Table]

	geo_v1_which	geo_v1_utm_zn	geo_v1_utm_esl
▶ Buena Vista	1	15	652559
Chimal	1	15	660133
Chichim	1	15	654875
Chuiguarabal	2		
El Cerro	1	15	644906
El Rancho	2		
La Ventosa	1	15	657115
Las Ventanas	1	15	648924
Los Chales	2	15	

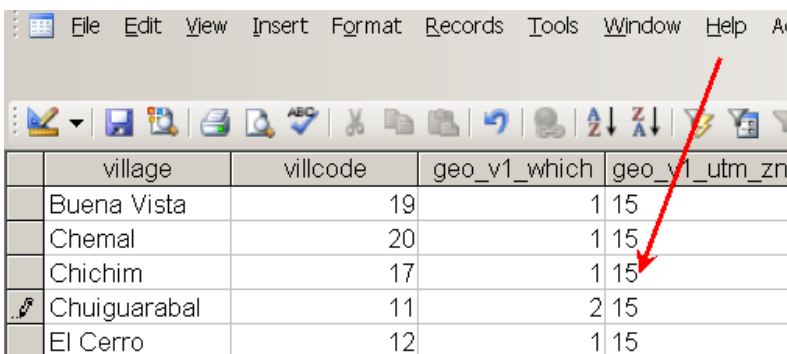
6. In the find box, enter the 11



7. Scroll to the relevant field



8. If the missing data is present in the questionnaire, then enter it in the database

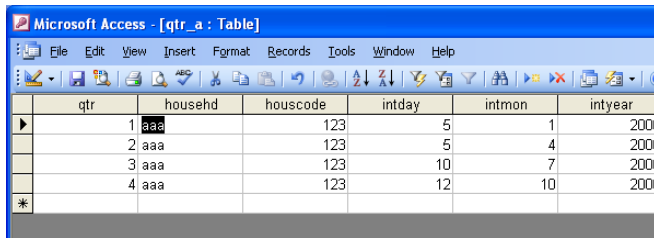


9. Repeat for all missing data for this table. When you are done, hit the save button to **SAVE**

Note:

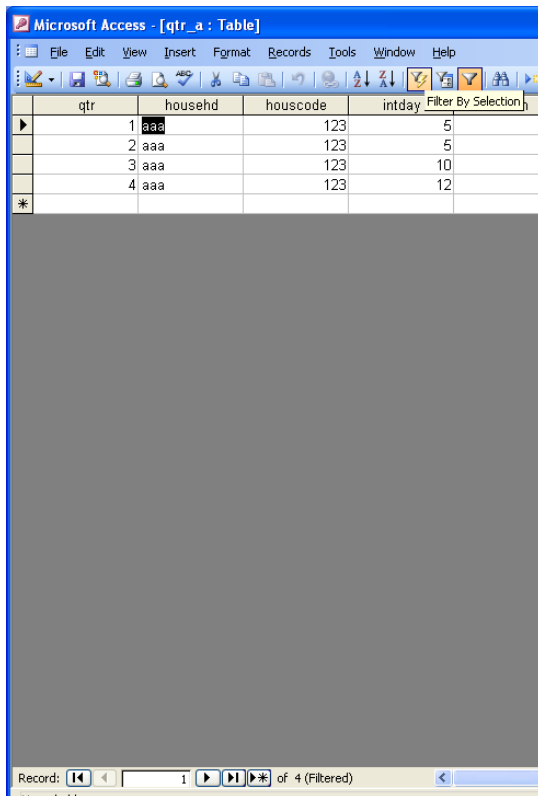
If the bug report lists data that are not unique on one key identifier (e.g. in the quarterly survey both household code and quarter are necessary to uniquely identify a record) after using “find” to locate the household code, apply a filter. Follow the steps below:

1. From the “Objects” menu on the extreme left of your screen, select “tables”.
2. Open the table in which you are going to correct bugs by double clicking on it and find the household code using steps 4-9 described above. On finding the household code, highlight it.



	qtr	househd	houscode	intday	intmon	intyear
▶	1	aaa	123	5	1	2006
	2	aaa	123	5	4	2006
	3	aaa	123	10	7	2006
	4	aaa	123	12	10	2006
*						

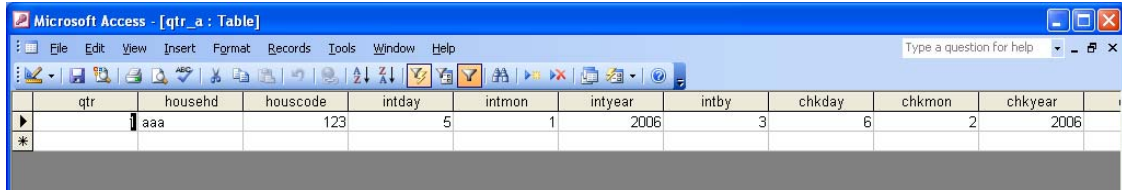
3. Click on the icon “filter by selection”. Note the number of records at the bottom of the screen with the words “filtered” in brackets.



	qtr	househd	houscode	intday	Filter By Selection
▶	1	aaa	123	5	
	2	aaa	123	5	
	3	aaa	123	10	
	4	aaa	123	12	
*					

Record: 1 of 4 (Filtered)

4. Highlight the number for the quarter with a problem and once again click the “filter by selection” icon.



5. Make the necessary corrections. To go back to looking at all the data in the table, click on the “remove filter” icon.

