THE RADON ANALYZER, A NEW DATA ANALYSIS TOOL

Robert K. Lewis
PA DEP, Bureau of Radiation Protection, Radon Division
Harrisburg, PA USA

INTRODUCTION

Since 1989 the Bureau of Radiation Protection's, Radon Division, has required by law the submission from the certified community, of both testing and mitigation data. The testing data submitted includes: name, address, postal city, zip code, county of test site, measurement location and result(s), measurement dates, measurement method, and house type. This data is either downloaded from disk or manually entered into an Oracle database on the mainframe computer. Mitigation data is also entered, but is not part of this analysis tool. There are approximately 1,000,000 test results encompassing about 666,000 locations.

In September 2003 the Bureau of Information Technology began working on this fourth in a series of "analyzers" for the Radon Division. The analyzer is based on Statistical Analytical Systems Institute (SAS) software.

Data Source Extraction & Transformation

Over 950,000 test result and location records from 1990 to 2003 were extracted from an Oracle database. Two Units of Measure (Working Level and Becquerel Meters Cubed) were excluded and only results in Pico Curies per liter were used. Extensive data validation involved comparing zip codes to appropriate counties, using the U.S. Post Office Global Zip Code Table and insuring that exposure times (test start and stop dates) were accurate for the specific test method. Data validation was made much easier in the Analyzer since it allowed for a much closer and more detailed analysis of individual records. The Analyzer data can also be exported to an Excel spreadsheet that also allows for further data manipulation and examination. Approximately 95,000 records failed the editing and were excluded. New data regarding DEP Field Units, Minor Civil Divisions, and Watersheds was added by an estimation process called geo-coding that designates a lat/long point by using street address and zip code. Ultimately, this yielded a new file of about 855,000 records, grouped into 48 fields with 544 characters per record.

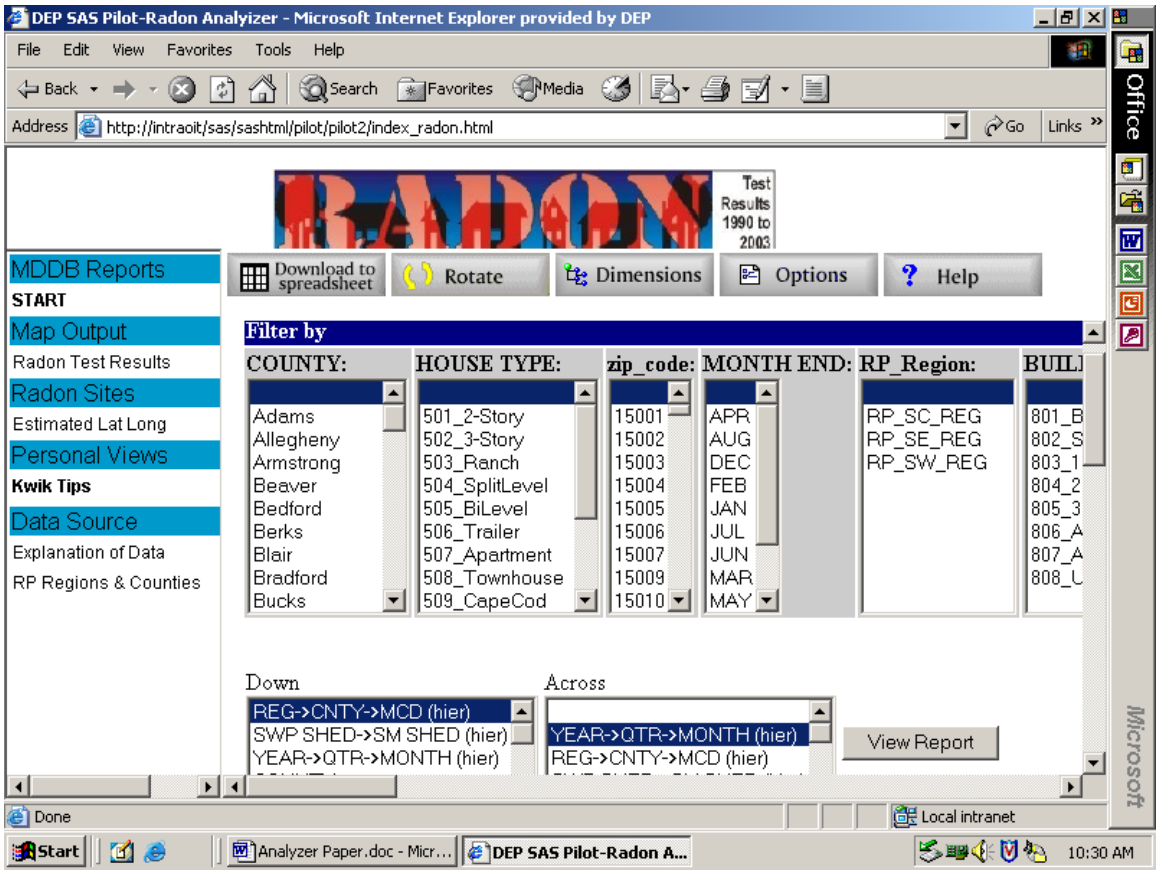
Environmental Analyzer

The Radon input file was loaded to the Analyzer, a special research web-enabled database made by SAS Institute. This tool pre-summarizes all those records having certain predefined, logical categories such as a zip code or test values above 100. This pre-summarized data all appears in the "down" and "across" columns. In total there are 16 pre-summarized categories of data. The data under the "filter by" columns is not pre-summarized. The tool also builds hierarchies of "nested" categories (such as

municipalities within counties) or unrelated “families” of data (such as zip codes crossing counties). At the highest level it provides a coherent multi-dimensional view of all the data. (It is technically known as a Multi-dimensional Database or MDDB and sometimes referred to as a “Cube.”) The Analyzer is able to bring responses to queries and reports on all logical categories within about 8 seconds to any user’s desktop that has a browser and is connected to the internal Local Area Network. The tool also provides 21 different statistical routines, color graphs, and downloads to spreadsheets.

The Radon Analyzer

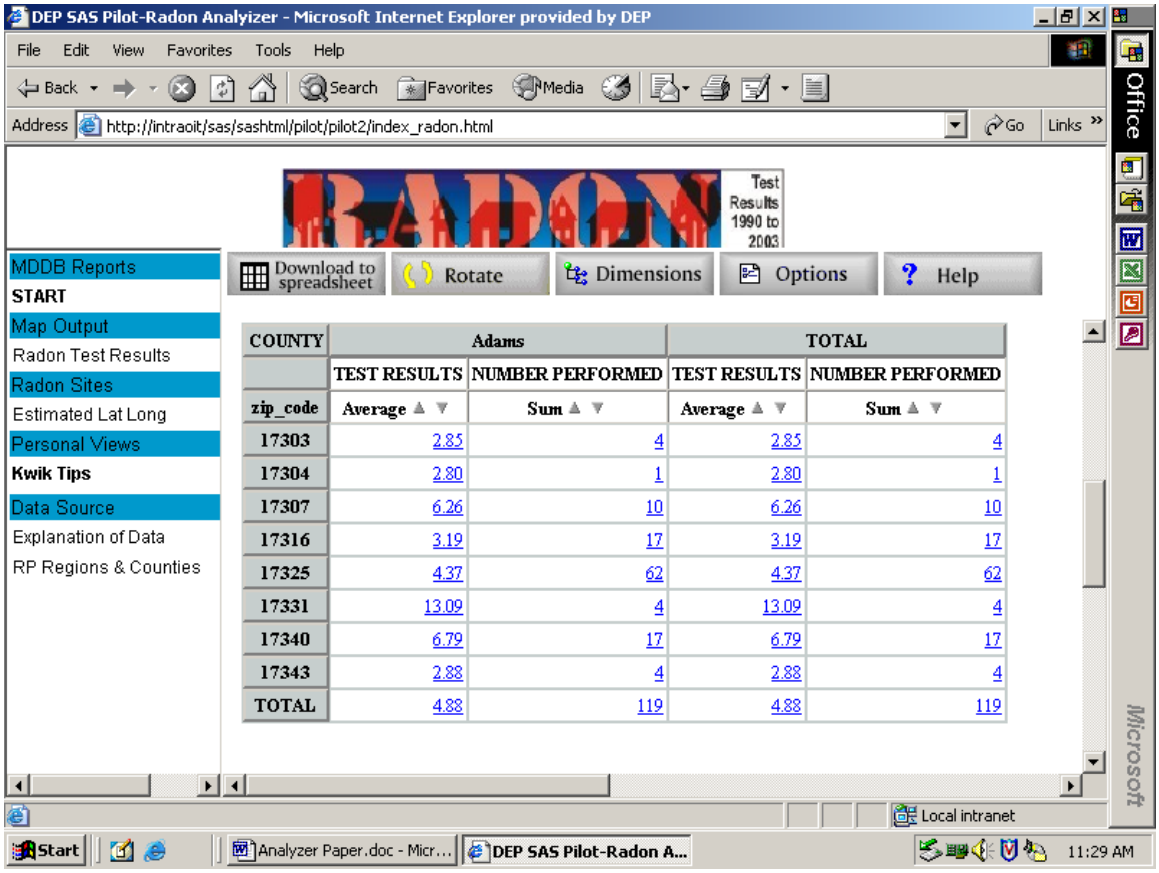
The diagram below shows the initial working screen of the Analyzer.



The list of items on the far left are static and for information only. They produce a Pennsylvania map showing county average radon concentrations and Kwik Tips provides information on menus, downloads, printing, and personal views. The Explanation of Data explains how and why the radon records were selected, derived, transformed, and loaded to the Analyzer. The five top row buttons provide for additional support and analysis. The Download to spreadsheet and Help buttons are self-explanatory. The Rotate button rotates the X and Y-axis on a graph (not visible in the above diagram). The Dimensions button allows for 21 statistics to be applied filtered data. Examples of some of the statistics are: coefficient of variation, standard deviation, minimum, maximum,

range, and sum. The Options button gives optional settings for the filter list box, report, table, and graph.

Listed below is an example query using County = Adams, Zip Codes, Building location = Basement results, Measurement type = continuous radon, charcoal, liquid scintillation, and short-term EIC, and year = 1995.



The Analyzer gives the average radon result for each zip code in the county and the number of results for that zip code for the year 1995. The Analyzer can continue to give further valuable information. As can be seen above, both the average and the sum are underlined. That means that you can "drill" down further on the data and look at individual records such as: measurement location, house type, lat./long., exposure dates, etc.

Results

Below are four tables and two graphs to give you some idea of the data generated from the Analyzer and its capabilities.

Table 1

Location	Sample Size	Average Result (pCi/L)
Basement	629,809	7.18
First Floor	163,055	3.61
Second Floor	31,908	2.83
Third Floor	1,664	4.22
Slab-on-grade	46,779	4.75
Above Crawl Space	6,653	2.61

Qualifications: Years 1990-2003, all house types, all measurement types, all PA counties.
Arithmetic average.

Table 1 shows the average radon concentrations by level in a house. The basement, as would be expected, is highest and in conjunction with the first floor value shows the often-quoted first floor/basement ratio of 0.5. Slab-on-grade shows the second highest concentration, which would not be unexpected. The third floor value of 4.22 pCi/L is unusual in that it is higher than both first and second floors.

Table II, Measurement Methods by Year

		Measurement Methods					
Year	**All Msmt. Types**	**CR**	**AT**	**AC**	**LS**	**EL**	**ES**
1990	33,351	1,771	613	22,365	---	40	8,562
1991	37,308	1,795	4,127	18,239	4	114	13,029
1992	41,666	2,273	2,711	19,103	---	266	17,313
1993	43,956	3,260	2,250	18,177	9	394	19,866
1994	48,179	2,526	1,404	24,828	8	444	18,969
1995	57,744	2,988	731	29,640	3	356	24,026
1996	64,138	4,998	1,014	26,917	51	457	30,701
1997	65,600	6,199	838	22,582	2,253	521	33,207
1998	81,666	6,006	844	34,778	2,225	511	37,302
1999	75,026	10,754	1,108	24,184	2,510	471	35,999
2000	74,091	13,129	1,212	20,174	7,041	347	32,188
2001	75,988	14,195	967	21,666	7,745	271	31,144
2002	84,473	16,210	828	22,709	8,885	274	35,567
2003	75,416	17,071	1,027	21,991	10,342	150	24,835
Total	858,602	103,175	19,674	327,353	41,076	4,616	362,708

Table II shows the use of different measurement methods over time starting in 1990. Several interesting observations are that charcoal liquid scintillation measurements are

almost nonexistent from 1990 to 1995 and then take a jump in 1997. In 1990 activated charcoal is the dominant testing method, with short-term electret ion chambers taking the lead in 1996 and then holding the lead. For some reason 2002 showed the largest amount of testing per year in Pennsylvania with 84,473 tests.

Table III, County Data, 13 of 67 counties

County	Sample Size	Bsmt. Average pCi/L	Sample Size	1st Fl. Average pCi/L
Adams	2840	6.77	525	2.97
Allegheny	74367	5.66	11358	3.57
Armstrong	735	9.01	219	6.61
Beaver	5470	7.85	548	5.81
Bedford	797	10.38	164	4.1
Berks	18284	11.51	2788	7.11
Blair	4542	6.55	526	3.2
Bradford	1577	8.35	249	4.68
Bucks	63497	5.41	25597	2.78
Butler	9102	8.06	1209	5.48
Cambria	3707	6.76	394	4.69
Cameron	57	18.34	10	6.85
Carbon	1425	11.48	461	3.96

Table III shows county data for 13 of our 67 counties. It shows the basement and first floor averages and their respective sample sizes. This entire table in part allows the Radon Division to rank counties.

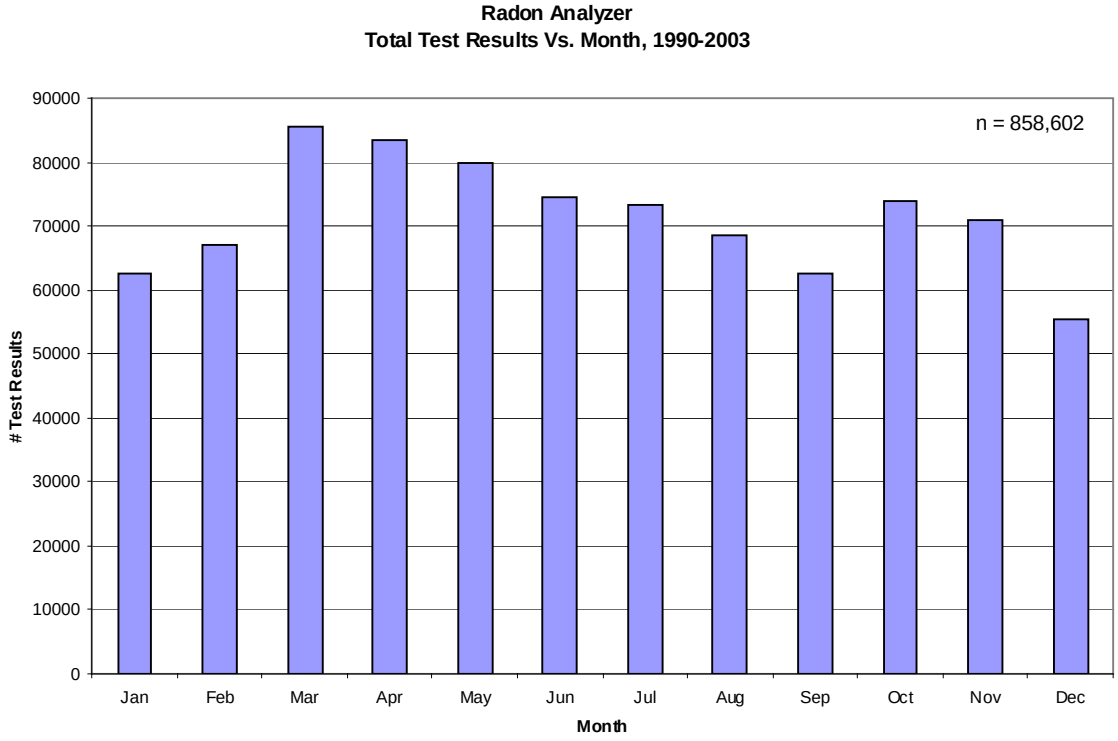
Table IV, Top Ten List of Counties

County	# Rslts > 100 pCi/L	Avg. of those >100	High Result (pCi/L)
Lancaster	201	161	631
York	190	167	957
Chester	186	207	1669
Berks	165	181	1866
Lehigh	164	162	879
Dauphin	119	160	918
Bucks	111	205	1126
Northampton	103	184	1400
Allegheny	90	137	325
Montgomery	70	211	803

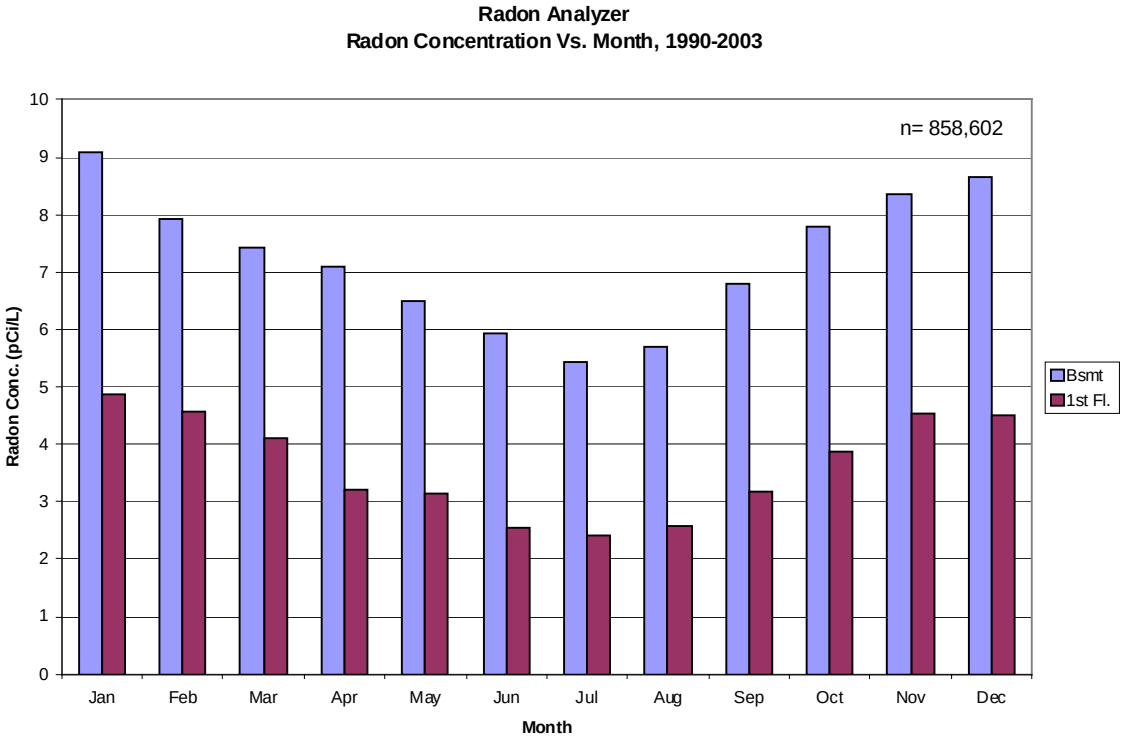
Qualifications: Years 1990-2003, all house types, measurement types AC, CR, ES, LS,

arithmetic average, duplicates excluded.

Table IV helps to locate some of our most problematic counties. All of these counties except for Allegheny are in the southeast part for the state, contiguous with one another, and running in a southwest to northeast orientation. Four of these counties also comprise our "Reading Prong", the "famous" physiographic province where the radon problem came to light.



This graph shows the distribution of testing by month. It would not be unexpected to see December as the lowest test time due to the Christmas, Hanukkah, and New Years celebrations. Why March and April show the highest testing is uncertain? Could it be tied to home sales?



This graph shows the seasonal distribution of radon concentration over time. It shows the typical winter highs (Jan., Feb., Nov., Dec.) and the summer lows (June, July, Aug.) for both basement and first floor.

Discussion/Conclusions

This new data analysis tool has provided a very powerful and user-friendly way to examine, sort, graph, and tally numerous radon testing data. With over 850,000 test results, a statistical result from any one query usually has a large sample size. Additionally, a data validation program by both the Bureau of Information Technology staff and Radon Division staff has helped to assure the quality of the data. One significant by-product of this data validation program was the discovery of numerous data errors in the original Oracle database. As a result, new more stringent edits are being put into place on the Oracle database, such as measurement units (pCi/L, WL, Bq/m^3) must be consistent with measurement method, exposure times for each measurement method are now specified within a range, zip codes must meet certain criteria, and zero results are converted to a minimal non-zero value.

Currently, this tool is only available to those within the Department. There are a number of problems in providing access to those outside of the Department. The Analyzer currently runs on Internet Explorer 6.0 and we do not know how it would run on another browser. Someone outside the Department would need a cable or DSL connection to get a good response time. A 56-K modem would take far too long to run queries. Finally, we have confidentiality issues with the data. We would have to at least remove the street address field to maintain individual test sites and to keep the results confidential.

This paper has only scratched the surface of what the Radon Analyzer can do. Hopefully, others will have access to this tool and find many additional uses of this new technology.

Acknowledgements

Special thanks to Mr. Ken Currie and Mr. Harit Trivedi of the Department of Environmental Protection, Bureau of Information Technology for their work in the development of the Radon Analyzer. Student intern, Nate Dysard also did much of the data validation work.