Upgrade of Illinois State Water Survey Groundwater Quality Database Year 1 Summary

Devin Mannix, Walt Kelly, Tom Holm, Greg Rogers

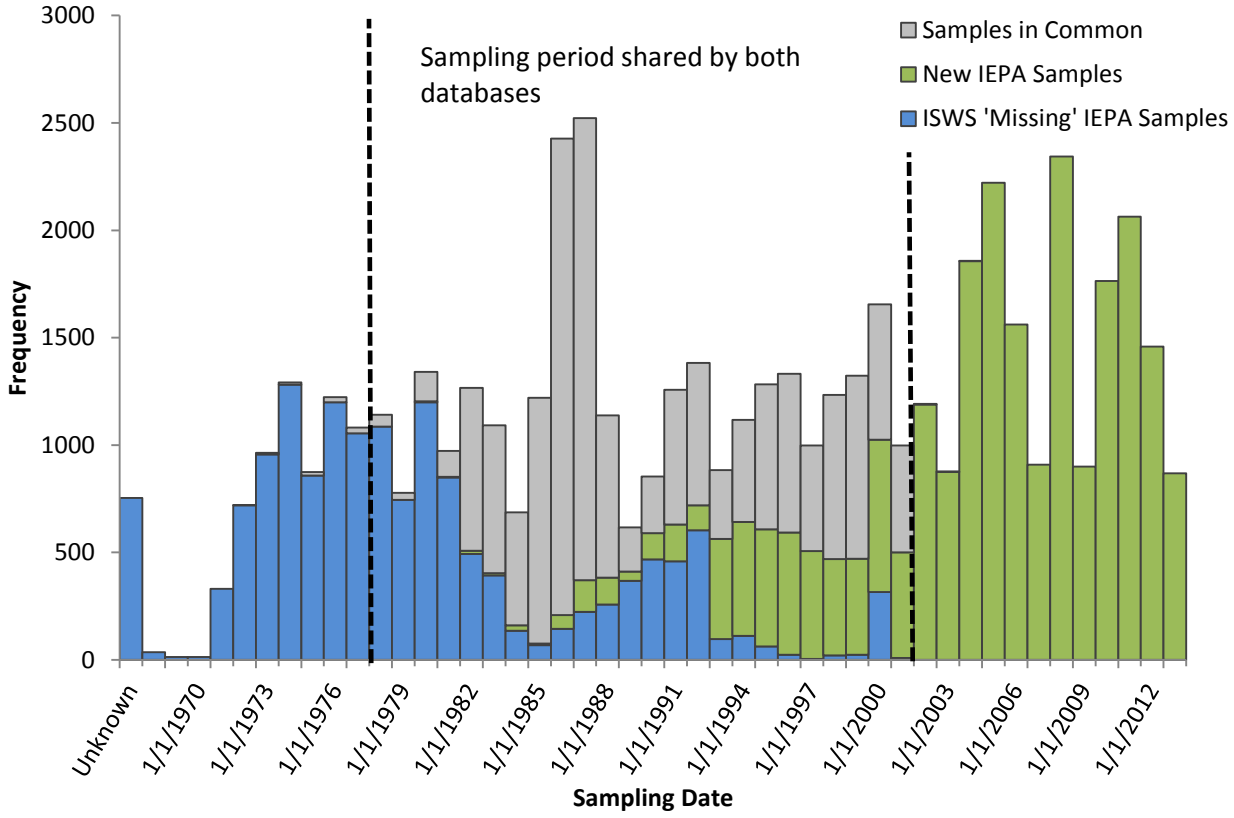
The Upgrade of Illinois State Water Survey Groundwater Quality Database project has six stated objectives:

Task 1: Get most recent copy of IEPA ambient water quality database
Task 2: Identify and correct errors in data
Task 3: Link water quality database with the IWIP and well databases
Task 4: Acquire scheduled updates of IEPA's ambient data
Task 5: Develop interface for online access of GWQDB
Task 6: Reconcile GWQDB and IEPA ambient network data discrepancies.

Database Merger

The project began after the acquisition of the most recent copy of the IEPA ambient groundwater database as of August 7, 2013. All samples were imported into the ISWS groundwater quality database structure. The combined database currently resides on a local machine, final import into the ISWS database pending. IEPA ambient groundwater samples already present in the ISWS database were filtered, leaving a combined total of 54,843 samples. Of these there are 15,551 samples unique to the ISWS database, though most of these fall outside the time period that the IEPA was actively sampling and represent inherited samples from other data sources (Figure 1). It is estimated that approximately 2,000 of these missing samples are part of the IEPA's sampling program. Furthermore, in the process of examining the original lab reports it became evident that there are a number of physical records in ISWS files that have not been entered into either database.

Figure 1. Sample discrepancies between ISWS groundwater quality database and IEPA ambient water quality database



In the process of merging the databases, a number of other issues were identified that highlight additional data discrepancies that will need to be addressed in the next phase of the project. One such issue involves the non-uniqueness of lab numbers. Care must be taken when assigning new sample IDs that a “sample” is identified by the lab number, well number, and collection date. Even so, there remain 1,436 samples in the ISWS database with duplicate records assigned to these samples, though sometimes the concentrations differ by a large margin. In a few cases this is due to a lab reporting multiple runs of the same sample, but in other cases it appears a different sample has been mistakenly entered or updated with the original sample’s information, rendering both unintelligible. Until the original lab reports are found and the records corrected in our database, these samples will remain unusable.

Well Identification

One of the problems encountered with IEPA legacy data stored in the ISWS database was the lack of unique well identifiers associated with some water quality samples. The ISWS Groundwater section has a database of approximately 425,000 wells in the state which are uniquely identified by their "p number." Additionally, for many PWS wells the IEPA well ID is also included along with other well metadata, such as PLSS coordinates, well depth, facility well numbers, municipal codes, and facility names. When the IEPA well ID is present in the groundwater quality database, the "p number" can be matched directly with the well database. However, for 11,805 samples, roughly half of the IEPA legacy data, the well ID is not known.

As a result, the only efficient means to identify the unknown wells was a query-based match ranking comparing the aforementioned fields between the two ISWS databases. This was an imperfect process, as evidenced by a pilot test using samples where the wells were known. Comparing well metadata between the well database and the groundwater quality database yielded a match percentage of 52.2% (Table 1), meaning almost half of the samples' location data no longer matched the well database in one or more fields. This can be traced back to data omissions and outdated well metadata present in the groundwater quality database. The first iteration of well matching had some obvious mismatches, largely because the facility has been shown to be incorrect for some samples. However, for the vast majority of samples the well matching appears to be accurate. Mismatches will become evident as the water chemistry is examined during the error checking process.

Table 1: Percentage agreement for ISWS location data parameters between water quality database

Criteria	Match Percent
Facility Name	**86.0**
Fuzzy Facility Name	92.4
FIPS	97.8
Township	99.2
Range	99.4
Section	96.0
Plot	77.9
Complete PLSS Match	**75.9**
Depth	**77.4**
Depth Within 10%	93.5
All Match	**52.2**

Data Quality

In order to assess the quality of the combined water quality database, a number of tests were performed at both the sample level and that of individual analytes. The first three tests, TDS error, conductivity error, and charge balance error, flagged any samples that did not fall within an acceptable margin of error, discussed later in this section. These tests required a minimum set of analytes to be present in the sample, which include Ca, Mg, Na, alkalinity, chloride, and sulfate. Additional analytes

considered, if present, include K, nitrate, ammonia, Fe, and SiO_2. Of the 54,843 samples, only 22,238 met the minimum requirements. To help fill this gap, the tests for outliers required only three measurements, at minimum, for any single analyte within a well's sampling history.

Unlike the first three tests, the tests for outliers do not flag an entire sample, rather they group all samples by analyte and well, then identify individual measurements that deviate from others of the same group. As the outlier tests are limited by an assumption of normality, an outlier may be indicative of a natural change in water chemistry, a misidentified well, a transcription error, or a lab error. The outlier tests are only meant to draw attention to these issues, especially when multiple outliers are present in a single sample.

The first outlier test was the Dixon's Q test, a basic test for flagging a single outlier within a sample set. A discussion of this test and critical values can be found in Bohrer (2008). The second test was a nonparametric test known as the modified Z-score, which is equivalent to the standard Z-score by means of an empirically derived constant, using the median and median absolute deviation to prevent extreme outliers from skewing the test statistic (NIST/SEMATECH, 2013). A modified Z-score of 3.5 was chosen to flag potential outliers. In a normal distribution this would account for approximately 99.95% of the dataset.

All solutions, including groundwater, are electrically neutral. This is the basis for calculating the charge balance error. For any samples that included the minimum analytes required, charge balance error was calculated by converting concentrations to milliequivalents, calculating the cation and anion sums and using equation 1:

Equation 1:
$$CBE = \frac{\Sigma cat - \Sigma an}{\Sigma cat + \Sigma an} \times 100\%$$

If a TDS measurement was included in the sample, TDS error was calculated from summing the aforementioned analytes in the sample and comparing it against measured TDS, following equation 2.

Equation 2:
$$TDSErr = \frac{TDS_{calc} - TDS_{meas}}{TDS_{meas}}$$

Similarly, if a conductivity measurement was present in the sample, calculated conductivity (Equation 5) was determined following the calculation of ionic strength (Equation 3), activity coefficients using the Davies approximation (Equation 4), then finally the conductivity error was calculated (Equation 6).

Equation 3:
$$I = \frac{\sum_i C_i z_i^2}{2}$$

Equation 4: $$Log(\gamma) = -0.5\left(\frac{\sqrt{I}}{1+\sqrt{I}} - 0.3I\right)$$

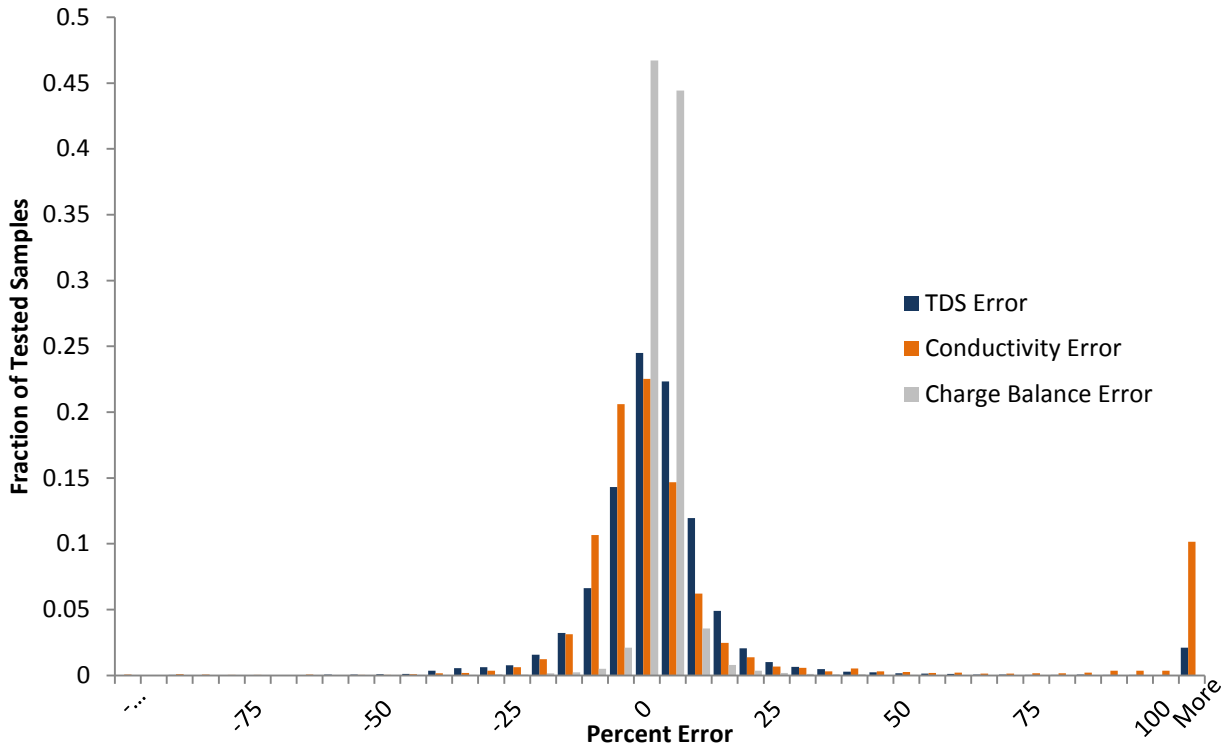
Equation 5: $$Cond_{calc} = \gamma^2 \sum_i C_i \lambda_i$$

Equation 6: $$CondErr = \frac{Cond_{calc} - Cond_{meas}}{Cond_{meas}}$$

Equation 5 is an empirical formula that has been used with some success in checking water analyses (Rossum 1975).

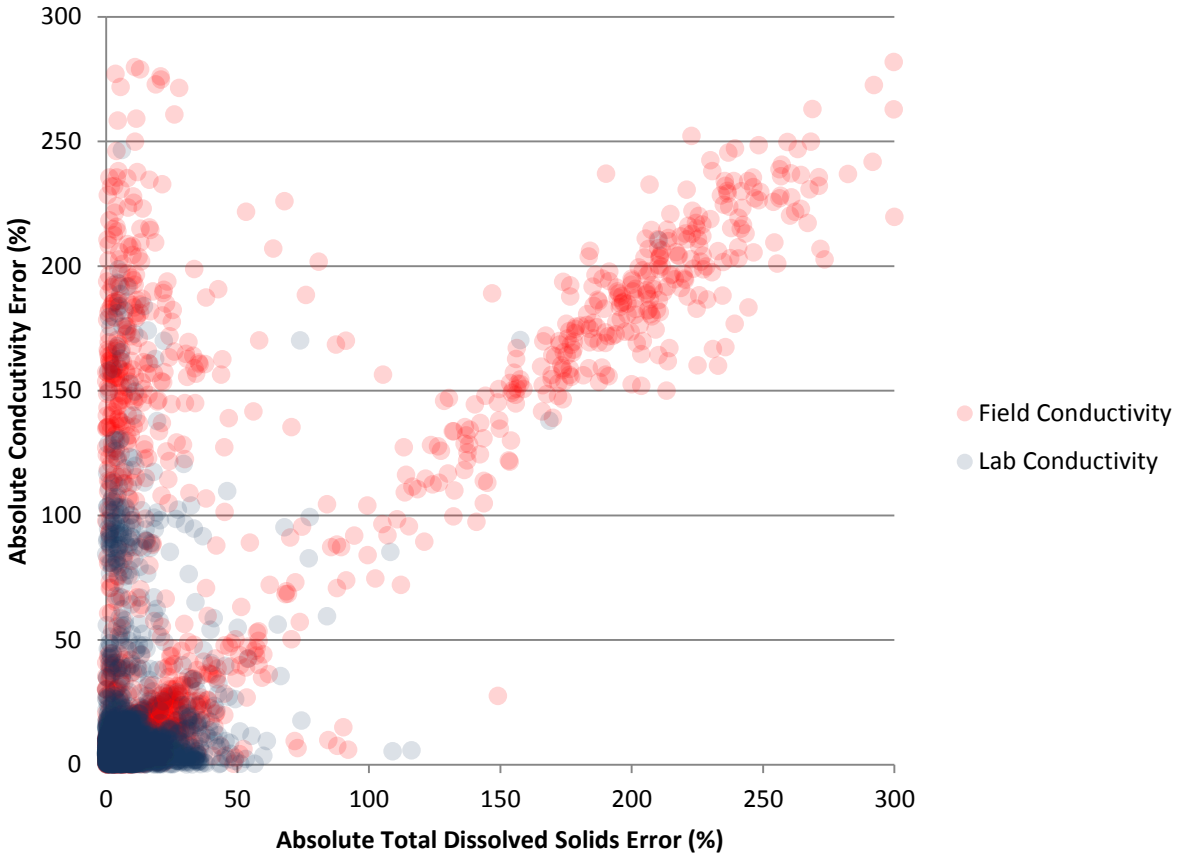
Figure 3 highlights the error distributions for these three tests. After examining the data, it became apparent that charge balance is tightly clustered around zero error for the vast majority of samples, so samples were flagged if the absolute charge balance error fell above ten percent. The error distributions for TDS and conductivity are broader and have some large positive errors in excess of 100%, so a tolerance of 20% error was chosen to flag samples using these tests.

Figure 3: Percentile distribution for TDS, Conductivity, and Charge Balance errors



It is interesting that the extreme errors seen in TDS and conductivity are often correlated, suggesting some of these “measurements” may be calculated values. In the IEPA legacy data, conductivity measurements are listed as either “field” or “lab” with the greatest errors associated with field values. This can be observed in Figure 4.

Figure 4: Correlation between TDS and Conductivity errors in IEPA legacy samples.



With these data quality checks in place, an Excel spreadsheet was developed using conditional formatting to highlight potential sample errors (Figure 5). There were three tiers of confidence assigned to these samples, which will be used in the final output if the identified problems cannot be corrected. Samples with the lowest confidence are those where the errors with TDS, conductivity, or charge balance exceeded their respective tolerances. In this case the entire sample was highlighted in red and the offending parameter was formatted to red and bold text. This flag is meant for samples that have obvious problems and should be avoided if possible. The middle tier involves the outlier checks; any measurement flagged as an outlier by at least one of the outlier tests was highlighted in yellow and formatted to bold text. As outlier checks are not always accurate for ambient groundwater samples, it is left to the end user to decide if the sample is acceptable, but when outliers are present the sample

should be treated with caution. Samples with the highest confidence are those that have passed these checks, particularly those where all of the tests could be performed. Using these criteria, a pilot test was performed on two counties checking for error patterns and potential solutions.

Figure 5. Example output for data quality checks. Each row represents a single sample.

	A	B	C	D	E	F	G	H	I	J	K	L	N	P	R	T	V	X	Z	AB	AD	AF	AH
1	sample_id	p_num	ISWS_Facility_ID	project_id	lab_num	date_collected	TDS	TDS_Calc	Cond	Cond_Calc	Cation Sum	Anion Sum	Ca reported (mg/L)	Mg reported (mg/L)	Na reported (mg/L)	K reported (mg/L)	Fe reported (µg/L)	NH4 reported (mg/L)	Alkalinity reported (mg/L)	Cl reported (mg/L)	SO4 reported (mg/L)	NO3 reported (mg/L)	SiO2 reported (mg/L)
146	49094	408524	01990250	156	B719588	12/16/1987	355	359		574.6	7.0	7.1	65	32	23	1.6	1062		350	1	10	0.1	16
147	59156	408524	01990250	156	B203165	3/10/1992	359	352	603	568.1	7.0	6.9	64.6	32.1	24.3	1.2	1100		338	1	10	0.00999	16
148	116093	411043	01990250	172	SF10264	6/3/2011	362	351	609	560.5	6.85	6.999	64.1	32.3	19.4	1.83	1110	1.15	350	0	0	0	16.8
149	116094	411803	01990250	172	SF10264	6/3/2011	346	347	605	557.9	7.034	6.699	64.6	32.6	22.7	2.01	1030	0.94499	335	0	0	0	17.2
150	64941	406982	01990300	156	B026772	2/24/1983	538	521	870	801.3	9.5	9.8	78	**23**	84	2.4	730		409	37	27	**0.1**	18
151	60541	406982	01990300	156	B401498	2/3/1994	**162**	503	**278**	782.3	9.4	9.3	74.9	28.6	75.2	1.9	739		387	32.8	32	0.00999	19.7
152	41570	406982	01990300	156	B501434	2/3/1995	520	584	875	884.2	10.1	11.7	82	31	77	2.1	870		507	36.4	27		18.3
153	41571	406982	01990300	156	B601802	2/5/1996	**244**	553	**397**	845.5	9.7	10.9	75	29	81	2.1	**1000**		463	37.7	28.1	0.00999	16.8
154	66419	406982	01990300	156	B900825	1/26/1999	487	555	**313**	819.7	9.689	10.54	75	29	80	2.1	750		428	37.8	26.6	10	19.9
155	110940	406982	01990300	172	GWB00227	2/17/2000	524	508	**363**	772.0	8.796	9.925	**65**	**25**	76	1.4	580	2.4	416	33.2	32.2	0	18
156	100334	406982	01990300	172	C6B0591	2/8/2006	593	501	861	781.0	9.117	9.43	78	29	61	1.9	790	1.97	371	39.8	42.7	0	18.4
157	84634	406982	01990300	172	08020714	2/15/2008	506	486	858	758.2	8.941	8.975	74.4	29.6	60.2	**3.2**	729	1.2	350	28	57	0	16.9
158	115065	406982	01990300	172	SA00977	1/29/2010	464	518	860	805.7	9.524	9.62	79.8	32	62.9	2	869	1.65	375	38.9	49.2	0	19.9
159	65031	406993	01990300	156	B032117	1/5/1981	436	441	740	684.8	8.2	8.3	62	26	67	1.8	370		369	19	20	0.1	18
160	64134	407004	01990300	156	B006698	5/8/1990	474	475	775	746.2	9.3	8.7	75.2	28.6	71	3.45	50		378	27	20	0.1	18
161	115180	410302	01990300	172	SA21018	1/27/2012	416	412	745	661.7	8.447	7.65	73.7	30.8	47.4	1.87	991	1.63	350	23.1	0	0	17.9
162	49438	402643	01990350	156	B050305	6/4/1978	370	379		589.6	7.3	7.2	55	26	52	2.1	1300	1.8	345	7.8	3.5	0	18
163	70066	402643	01990350	156	B711515	7/15/1987	408	386	**580**	607.0	7.544	7.304	62	26	51	1.8	1300		350	7	10	**0.1**	17
164	64169	402643	01990350	156	B010201	7/10/1990	375	392	627	577.3	6.5	7.8	53.5	23.6	41.7	1	1100		345	**16**	10	10	17
165	70067	402643	01990350	156	B010201	7/10/1990	375	370	627	577.3	6.477	7.454	53.5	23.6	41.7	1	1100		345	**16**	10	0.00999	17
166	100656	402643	01990350	172	C6G2288	7/21/2006	417	379	628	587.0	7.11	7.373	59	25	45	1.6	1200	1.22	359	6.9	0	0	18.3
167	114148	402643	01990350	172	S810467	7/31/2008	322	347	625	583.1	7.265	7.012	60.7	25.8	46	1.45		1.36	340	7.56	0	0	
168	116226	402643	01990350	172	SG00945	7/15/2010	378	391	624	603.6	7.126	7.646	61	26.4	40.2	2.28	1100	1.16	360	7.22	11.7	0	18.5
169	116345	402643	01990350	172	SG21567	7/24/2012	**304**	381	**641**	585.6	7.017	7.477	56.3	25.1	45.6	1.4	1220	1.44	365	6.34	0	0	19.6

Case Study: Champaign County

There are 544 records for Champaign County that contain the minimum analytes required for data quality tests in the combined database. There were a few missing values for TDS, nitrate, silica, Fe, and K. There were only 247 entries for conductivity and 288 for NH_3-N. Measured and calculated values of TDS and conductivity agreed within 20% for over 90% of the samples. Over 98% of the records had charge balance errors less than 10% (Table 2).

There were about half as many outliers for conductivity as for TDS (Table 2). However, there were also only about half as many values. Similarly, there were relatively few NH_3-N outliers, but also relatively few NH_3-N values. Silica concentrations varied over a narrow range, so it is not surprising that there are so few outliers. Nitrate had the most outliers, possibly because of many values below detection (Table 3).

Table 2. Statistics for TDS, conductivity, and charge balance for Champaign County samples.

Measurement	Records with Data	Outliers	Error Criteria (RPD[a])	Number of Records with Errors[b]
TDS	540	29	20	47
Conductivity	247	16	20	50
Charge Balance Error	--	--	10	11

Notes: [a]Relative percent difference

[b]Measured vs calculated values for TDS and conductivity. Anions vs cations for charge balance.

Table 3. Statistics for chemical analyses for Champaign County samples.

Measurement	Records with Data	Outliers	Measurement	Records with data	Outliers
Ca	544	18	Alkalinity	544	28
Mg	544	33	Chloride	544	33
Na	544	25	Sulfate	544	28
K	533	30	Nitrate	542	45
Fe	541	38	Silica	531	12
NH_3-N	288	7			

Many IEPA lab reports corresponding to database records were found in the Champaign County files. Nearly all reports had values for all measurements. Therefore, we will be able to correct some records for missing data. Several reports were found for which there are no records in the combined database. We will add these records to the database.

Comparison of the database records with lab reports found remarkably few transcription errors. Some of these errors were found by inspecting the spreadsheet for charge balance errors and disagreement between measured and calculated TDS and conductivity values. For example, Figure 6 shows a screen shot of the records for Mahomet. The spreadsheet highlighted the questionable TDS value in red. The actual value on the lab report was 519 mg/L. Figure 1 also shows that several records are missing conductivity or NH_3-N values.

Figure 6. Screen shot of database QA spreadsheet showing data for Mahomet (anions not shown).

sample_id	p_num	ISWS_Facility_ID	project_id	lab_num	date_collected	TDS	TDS_Calc	Cond	Cond_Calc	Cation Sum	Anion Sum	Ca reported (mg/L)	Mg reported (mg/L)	Na reported (mg/L)	K reported (mg/L)	Fe reported (µg/L)	NH4 reported (mg/L)
49220	408245	01990450	156	0021659	1/1/2050		549		913.6	10.8	10.7	107	56	18		1500	
49221	408234	01990450	156	0021660	1/1/2050		507		849.4	10.0	9.9	102	50	17		1800	
49214	408256	01990450	156	B001245	7/5/1972	394	384		607.5	7.4	7.2	78.8	31.9	19	1	400	0
49222	408234	01990450	156	B001246	7/5/1972	537	504		816.9	9.5	9.4	101.2	48	9	2	500	0
49219	408256	01990450	156	B026611	2/24/1983	434	412		642.8	7.6	7.8	83	29	24	1.7	910	
49211	408256	01990450	156	B026811	2/24/1983	434	412	650	642.8	7.6	7.8	83	29	24	1.7	910	
49226	408234	01990450	156	B031412	12/29/1980	517	505		816.3	9.6	9.4	103	46	12	2.4	2160	
49218	408256	01990450	156	B031413	12/29/1980	398	413		650.2	7.8	7.8	85	32	20	1.7	1050	
49224	408234	01990450	156	B049200	6/14/1976	470	524		838.4	9.7	9.9	**102**	48	11	2.2	2000	0.72
49216	408256	01990450	156	B049202	6/14/1976	**350**	402		630.8	7.6	7.8	**79**	32	19	1.4	1400	1.32
49225	408234	01990450	156	B050774	6/6/1978	505	500		802.9	9.3	9.4	91	49	13	2.5	**2000**	**0.8**
49217	408256	01990450	156	B050775	6/6/1978	398	398		624.4	7.6	7.5	74	34	21	1.8	1400	1.8
49223	408234	01990450	156	B112130	5/14/1974	**6**	493		795.2	9.3	9.2	96	46	12	2	2180	0.6
49215	408256	01990450	156	B112131	5/14/1974	367	400		624.0	7.5	7.7	80	30	19	1.5	1420	1.6
115835	412251	01990450	172	SD11293	4/28/2011	328	391	662	616.9	7.817	7.065	81.6	31.1	24.1	1.63	1490	0.781

In some cases bad TDS or conductivity values were confirmed by inspection. For example, Figure 7 shows a screen shot of the records for Young's Hillcrest Mobile Home Park. Only the sample information and major cation concentrations are shown. For each metal, the concentrations cover a narrow range, whereas two of the TDS values are quite different from the others and were flagged by the spreadsheet.

Figure 7. Screen shot of database QA spreadsheet showing data for Young's Hillcrest Mobile Home Park (anions not shown).

sample_id	p_num	ISWS_Facility_ID	project_id	lab_num	date_collected	TDS	TDS_Calc	Cond	Cond_Calc	Cation Sum	Anion Sum	Ca reported (mg/L)	Mg reported (mg/L)	Na reported (mg/L)	K reported (mg/L)
49285	405020	01990040	156	0024285	1/1/2050	350	379		599.4	7.5	7.4	54	27	56	
62789	405021	01990040	156	0024286	1/1/2050	340	377		610.2	7.5	7.3	54	27	56	
49186	405021	01990040	156	B051017	6/21/1976	310	348		550.0	6.9	6.7	52	28	42	2.2
49288	405020	01990040	156	B051018	6/21/1976	**283**	371		577.7	7.1	7.3	56	28	42	2.1
49289	405020	01990040	156	B053110	6/14/1978	356	364		570.9	7.3	6.9	57	29	42	2.2
49187	405021	01990040	156	B053112	6/14/1978	343	371		579.0	7.2	7.2	56	29	43	2.4
49286	405020	01990040	156	B101952	8/23/1973	**503**	366		569.2	7.1	7.0	54	27	46.7	2.6
49185	405021	01990040	156	B110636	4/16/1974	381	358		563.2	7.1	6.9	53	28	45	2.2
49287	405020	01990040	156	B113505	5/23/1974	396	361		564.2	7.1	6.9	52	29	45	2.4

Figure 8 (Thomasboro) shows how a disagreement between measured and calculated TDS combined with a bad charge balance led to the discovery of a questionable Na concentration. The Na value for the highlighted records is probably 22.4 and not 224.

Figure 8. Screen shot of database QA spreadsheet showing data for Thomasboro (anions not shown).

sample_id	p_num	ISWS_Facility_ID	project_id	lab_num	date_collected	TDS	TDS_Calc	Cond	Cond_Calc	Cation Sum	Anion Sum	Ca reported (mg/L)	Mg reported (mg/L)	Na reported (mg/L)	K reported (mg/L)
49252	408479	01990950	156	B001253	7/10/1972	354	344		546.3	6.8	6.7	61.2	31.9	24	1.3
64971	408468	01990950	156	B027686	3/2/1983	338	350	580	556.9	6.7	6.9	60	31.7	**23**	2.1
64972	408479	01990950	156	B027687	3/2/1983	326	352	590	559.5	6.8	6.9	60	32	24	2.1
62788	408468	01990950	156	B033864	1/13/1981	318	351		560.0	6.9	6.8	63	31.9	23	1.8
49257	408479	01990950	156	B033867	1/13/1981	318	352		563.2	6.9	6.9	63	32	24	1.6
49250	408468	01990950	156	B051994	6/29/1976	407	365		587.3	7.0	7.2	64	32	21	1.7
49255	408479	01990950	156	B051996	6/29/1976	335	351		554.2	6.8	7.0	60	31	23	1.7
49256	408479	01990950	156	B053115	6/14/1978	344	350		557.4	7.0	6.8	63	32	23	2
49251	408468	01990950	156	B053117	6/14/1978	320	343		546.5	6.8	6.6	62	**32**	21	2
70105	408491	01990950	156	B101099	1/24/1991	344	355	561	564.9	6.89	6.917	64.5	**32.1**	**21.4**	1.8
49247	408468	01990950	156	B104476	10/31/1973	**386**	**553**		889.9	**15.8**	**6.9**	61	34	224	1.9
49254	408479	01990950	156	B115048	6/18/1974	359	348		554.6	6.9	6.7	64	30	25	1.6
49249	408468	01990950	156	B115050	6/18/1974	**347**	359		567.9	6.9	7.1	64	31	**23**	1.7

One of the records for Broadlands was flagged for a suspect TDS value (Figure 9). Inspection of the data and comparison with other Broadlands records confirmed that the suspect TDS value was probably in error. Further inspection of the Broadlands records revealed that two sub-groups of Broadlands records are chemically distinct. There are 13 records in the combined database, indicated in Table 4 by rows with entries in the P Number (ISWS well ID number) column. IEPA lab reports were found for five of the records and two more reports were found for which there are no entries in the database. These are indicated by rows with entries in the Well column and blank cells in the P Number column. The Mann-Whitney U test was used to compare the median Ca, Mg, Na, K, alkalinity, chloride, and sulfate values for the known Broadlands records and the other records (those with blank cells in the Well column). The results are in the last row of Table 4. The median values for the two sub-groups were found to be significantly different with a confidence level better than 0.01 for 5 of the 7 measurements. The last 8 records in Table 4 may be for another facility. We will try to find out which facility these records really belong to.

Figure 9. Screen shot of database QA spreadsheet showing data for Broadlands (anions not shown).

sample_id	p_num	ISWS_Facility_ID	project_id	lab_num	date_collected	TDS	TDS_Calc	Cond	Cond_Calc	Cation Sum	Anion Sum	Ca reported (mg/L)	Mg reported (mg/L)	Na reported (mg/L)	re
49379	407137	01990050	156	0104305	11/20/1972	**341**	419		608.6	7.8	7.7	**52**	**25.5**	**67**	
49349	407137	01990050	156	B002186	7/17/1972	453	417		668.7	8.2	8.0	**85**	**36**	**21**	
49383	407148	01990050	156	B026775	2/24/1983	374	**406**		617.7	7.7	7.5	55	19.2	75	
49382	407148	01990050	156	B032120	1/6/1981	382	382		575.3	7.0	7.1	**49**	**18**	**69**	
62779	406971	01990050	156	B032876	1/13/1981	407	**436**		693.0	8.4	8.2	89	35.2	23	
62777	407137	01990050	156	B033878	1/13/1981	402	427		685.7	8.3	8.1	86	36.6	21	
62778	406971	01990050	156	B050653	6/21/1976	367	398		632.0	7.8	7.6	80	33	21	
49352	407137	01990050	156	B050661	6/21/1976	390	421		677.0	8.3	8.0	86	36	20	
49381	407137	01990050	156	B051022	6/22/1976	347	370		562.7	7.0	6.9	47	18	68	
49355	406971	01990050	156	B051223	6/6/1978	426	427		672.1	8.2	8.1	83	35	23	
49353	407137	01990050	156	B051224	6/6/1978	415	431		684.5	8.4	8.1	81	38	23	
49350	407137	01990050	156	B100940	7/30/1973	462	429		686.5	8.4	8.2	86	37	20	
49380	407137	01990050	156	B114517	5/6/1974	407	363		547.1	6.8	6.7	46	17	68	

Table 4. Major ion concentrations in Broadlands well water.

P Num	Sample ID Number	Date Collected	Well	Ca	Mg	Na	K	Alk	Cl	SO4
407137	B114517	5/6/1974	1	46.0	17.0	68.0	1.1	332.0	3.0	0.0
407137	B051022	6/22/1976	1	47.0	18.0	68.0	1.1	336.0	4.3	4.2
407148	B032120	1/6/1981	2	49.0	18.0	69.0	1.2	346.0	3.6	5.0
	B032845	3/12/1985	2	51.0	18.1	67.0	1.3	338.0	4.0	10.0
407137	0104305	11/20/1972	1	52.0	25.5	67.0	1.3	348.0	3.0	10.0
407148	B026775	2/24/1983	2	55.0	19.2	75.0	1.3	359.0	3.2	11.0
	B603675	3/15/1986	3	63.0	29.0	112.0	3.6	560.0	9.5	10.0
406971	B050653	6/21/1976		80.0	33.0	21.0	1.7	360.0	7.8	10.0
407137	B051224	6/6/1978		81.0	38.0	23.0	2.0	357.0	14.0	29.0
406971	B051223	6/6/1978		83.0	35.0	23.0	2.1	368.0	11.0	21.0
407137	B002186	7/17/1972		85.0	36.0	21.0	1.9	358.0	12.0	19.0
407137	B033878	1/13/1981		86.0	36.6	21.0	1.6	356.0	14.0	28.0
407137	B050661	6/21/1976		86.0	36.0	20.0	1.7	350.0	14.0	27.0
407137	B100940	7/30/1973		86.0	37.0	20.0	1.7	364.0	13.0	26.0
406971	B032876	1/13/1981		89.0	35.2	23.0	1.6	364.0	12.0	28.0
		Confidence level:		0.0014	0.0014	0.0013	0.0228	0.0726	0.0020	0.0035

Case Study: Kane County

There were a total of 806 samples from public water supplies in Kane County, and samples with potential errors were flagged using techniques described above. The numbers of records flagged for each specific test are reported in Table 5.

Table 5. Numbers of records flagged for specific errors.

Type of Error	Number of Records
Missing Facility Information	27
Missing Date	8
Ion Balance Error	25
Calculated TDS Error	91
TDS-Conductivity Error	77
Calcium out of range	37
Potential Outliers	
Magnesium	39
Sodium	43
Potassium	31
Alkalinity	52
Chloride	61
Sulfate	34
Nitrate	101
Silica	54

Paper records in the ISWS Groundwater Section Records Room were examined to determine if flagged samples could be corrected. We successfully corrected a number of errors (Table 6). Once again, examining the paper records revealed a significant number of water quality samples (97) that were not part of the electronic database. These will need to be entered into the database manually. Many of these missing records were from the 1960s and early 1970s, prior to IEPA's existence, mainly collected by the Illinois Department of Public Health. A large number, however, were from 1989, which was surprising. The ISWS Records Room generally does not have paper copies for samples collected after 1989. A duplicate lab number (B032676) was identified, with a sample from Aurora and North Aurora both having that number. Flagged records that could not be corrected were given a confidence ranking based on criteria discussed above.

Table 6. Errors fixed using information from paper records at ISWS.

Type of Error	Number of Records
Facility Identification	86
Sample date	8
Parameter concentration	38
Sample identified as "finished"	5
Incorrect Lab number	12

This kind of assessment needs to be done for the rest of the state. Most counties will have many fewer samples than Kane County, which is one of the largest users of groundwater in Illinois, though these case studies in Champaign and Kane counties reveal the types of problems we can expect to find with the remaining records. In particular, the discovery of samples that have not been entered into the water quality database will need to be rectified. A standard process is being developed for searching through the files, verifying the records are present in the database, making note of discrepancies, and correcting any errors with the sample.

Error Frequency

Plotting the flagged sample frequency over time using the selected criteria reveals some temporal trends. Though the number of samples meeting the minimum testing requirements has decreased sharply following the 1980s (Complete Test Samples in Figure 10), the error frequency has increased as a percentage of total samples tested (Figure 11).

Figure 10: IEPA sample error frequency

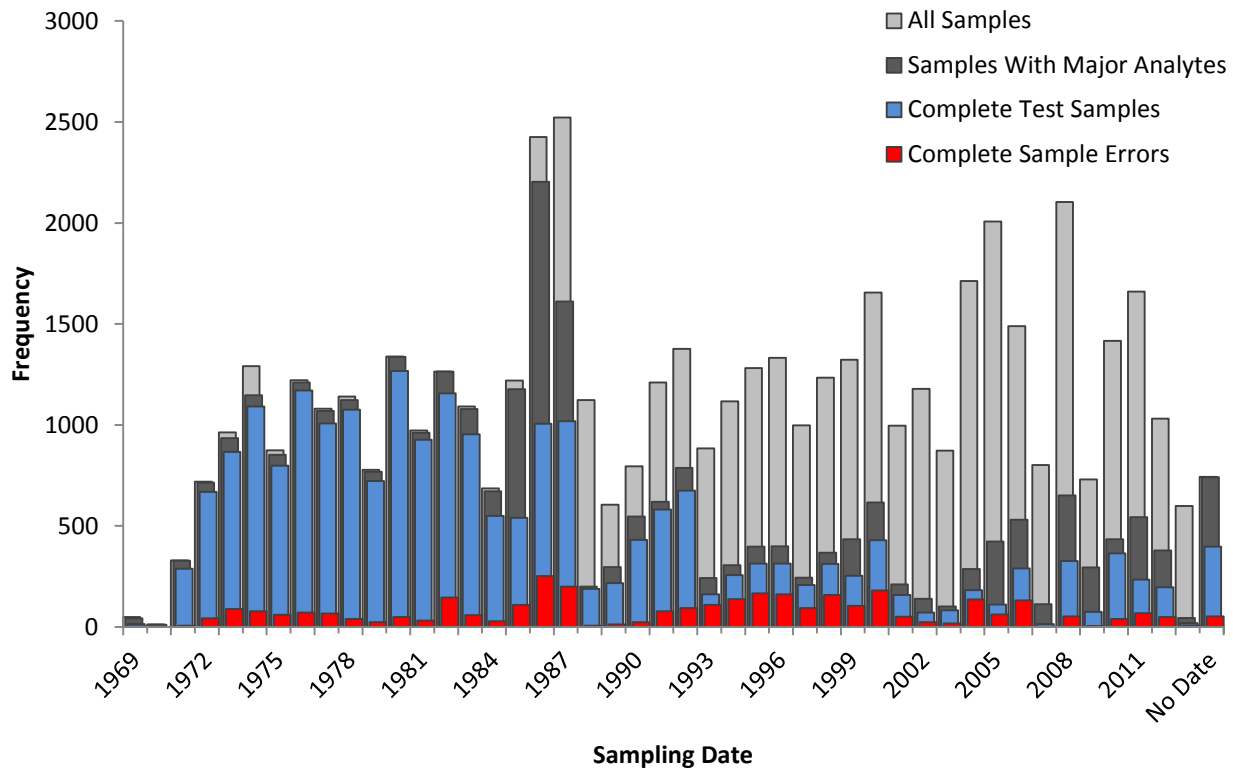
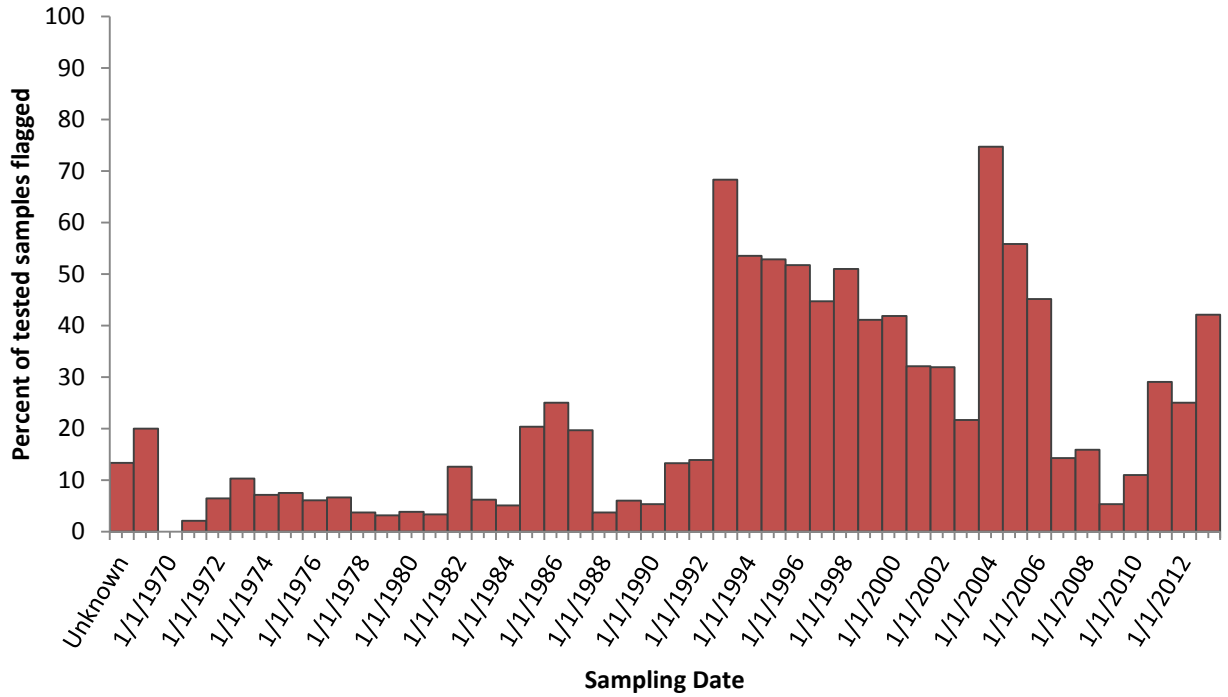


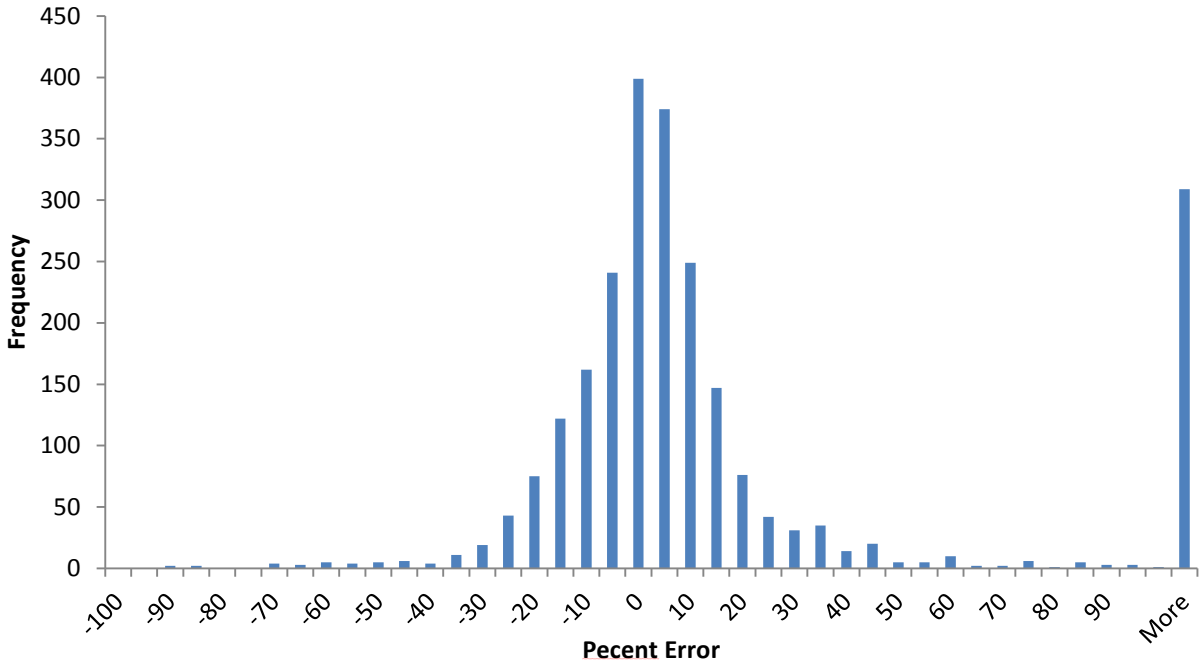
Figure 11: IEPA Sample error percentage by year



A few patterns emerge from this related to the labs during the periods in question. One of the more persistent problems is seen in samples with “D” lab numbers. Though these samples are predominantly composed of organics, when they do contain major inorganic analytes they are almost always flagged as outliers, as the sample appears drastically different from other samples from the same well. In contrast, TDS, conductivity, and charge balance are not often flagged. This problem is only seen in the IEPA legacy data as it appears recent copies of IEPA samples have already stripped “D” samples of these analytes. The same action will likely be taken with the legacy samples before the database is finalized.

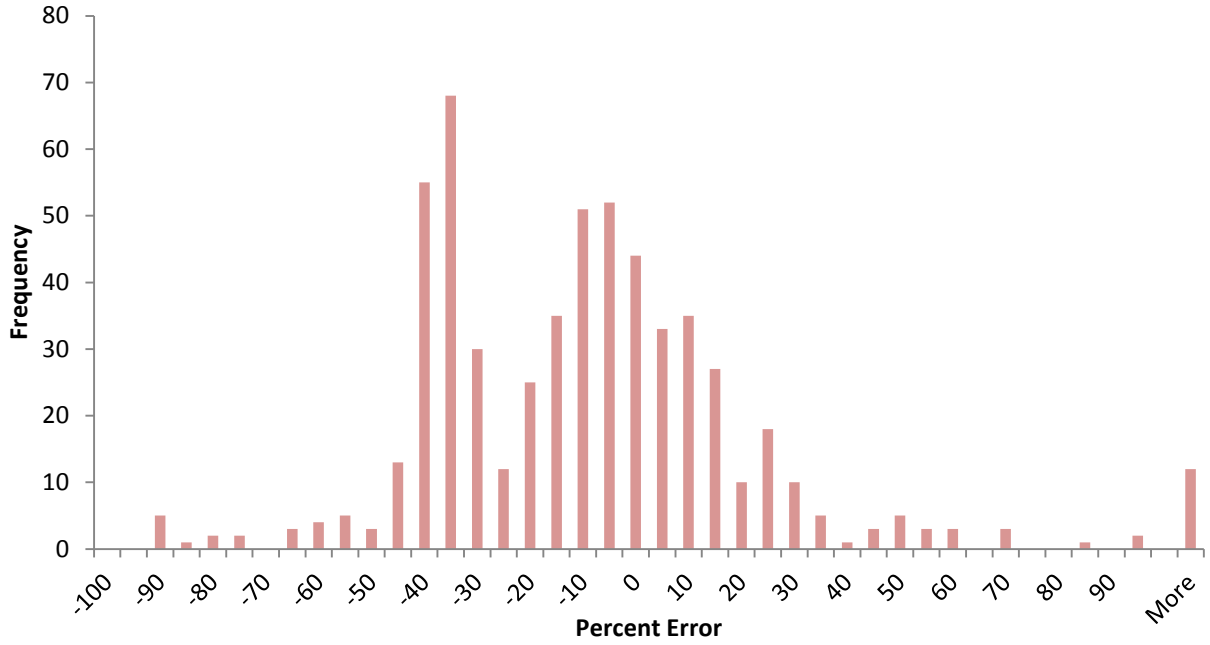
Another problem is seen in samples with “B” lab numbers. This is occurs over a specific period beginning approximately in 1993 and tapering off by the early 2000s, where either TDS is underreported or the individual analytes are erroneously high, as evidenced by the large positive errors (Figure 12). Conductivity measurements have similar problems, though there are fewer of these in the database.

Figure 12: “B” lab sample TDS error frequency for the period of 1993-2003



Directly at the tail end of this period follows frequent problems with the “C” lab samples from 2004-2006, again primarily flagged by TDS errors. The pattern is less consistent for this period, though the distribution is nearly bimodal with the larger peak in the negative percent errors (Figure 13).

Figure 13: “C” lab sample TDS error frequency for the period of 2004-2006



In the most recent record there appears to be another uptick in sample errors, though this may be a dubious trend as there are fewer samples for this period, and the errors don’t appear to be isolated to TDS. As we do not have paper records beyond the early 1990s, the most recent samples will be more difficult to verify, but the fact that errors persist highlights the continuing need for data quality checks.

Year 2 and Beyond

By the end of the first year of this project, we successfully identified a number of issues with IEPA groundwater quality data. At the onset of this project it was believed that reported TDS would be a reliable parameter in assessing data quality, but after examining the data it would seem that these measurements are often in error. Reported conductivity appears to be even less reliable, though conductivity has not been reported for many samples. These measurements should be used with extreme care. Charge balance, in contrast, remains acceptable for the vast majority of samples.

The process of error correction will proceed for the rest of the state and will require verifying samples with the paper records whenever possible. Additional data quality checks, primarily outlier tests, may be performed on minor analytes for both the original dataset and any future updates. When these corrections are finalized, all samples in the combined database will be returned to the IEPA including the confidence levels discussed earlier. We suspect most identified problems will be unresolvable, and other

problems may arise as the paper records are examined. In particular, we do not yet know the extent of records missing from our database. Once this phase of the project is completed, future updates should be scheduled on an annual basis. As the tools are already in place and there will be many fewer samples to examine, problems may be identified early in the process before they become systemic as we have seen in the past.

The long term vision with regards to the Ambient Water Quality data is to integrate it within a larger PWS portal where users will have access to water quality, water level, and other water usage data from water facilities across the state. Authorized users will have access to these data through multiple views and reporting tools. Researchers will have the ability to perform groundwater modeling and analysis using these data while stakeholders will have a central repository to easily monitor compliance and perform basic quality assurance functions. The general public will also be permitted to view these data on a reduced scale with sensitive or potentially risky data filtered out. Regular updates of the water quality data will be performed and made available on an annual basis to ensure the most current data are accessible. Additionally, a suite of web services will be provided to enable the dynamic exchange of data with other data systems, including SDWIS, the Federal Safe Drinking Water Information System.

References

Bohrer, 2008. One-side and Two-sided Critical Values for Dixon's Outlier Test for Sample Sizes up to n = 30. Economic Quality Control, v. 23 no. 1.

NIST/SEMATECH, 2013. e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm, June 20, 2014.

Rossum, J. R. 1975. Checking the Accuracy of Water Analyses through the use of Conductivity. J. Am. Water Works Assoc. 67:204-205.