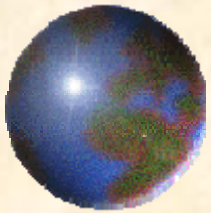


How Accurate Is Your Data?

2007 SERUG

Jacksonville, Florida

May 2-4, 2007

Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, Fl. 33410

How Accurate Is Your Data?

- "Good Enough"
- "Pretty Good"

- We often don't know until we compare it to another data set.
- Then the question is, "Which data set is the accurate one?"

Value of Data

- Value is less related to the cost of producing the data and more on its fitness for a particular analysis.
- Data quality significantly affects confidence in analysis results.
- Unknown data quality leads to tentative decisions and increased liability.

The 5 Components of Data Quality

- Positional Accuracy
- Attribute Accuracy
- Logical Consistency
- Completeness
- Lineage

Positional Accuracy

- How closely the coordinate description of features compare to their actual location.
  - Absolute Accuracy
  - Relative Accuracy

Attribute Accuracy

- How thoroughly and correctly the features in the data set are described.

Logical Consistency

- The extent to which geometric problems and mapping inconsistencies exist within a data set.

Completeness

- The decisions that determine what is contained in the data set.

Lineage

- What sources are used to construct the data set and what steps are taken to process the data.

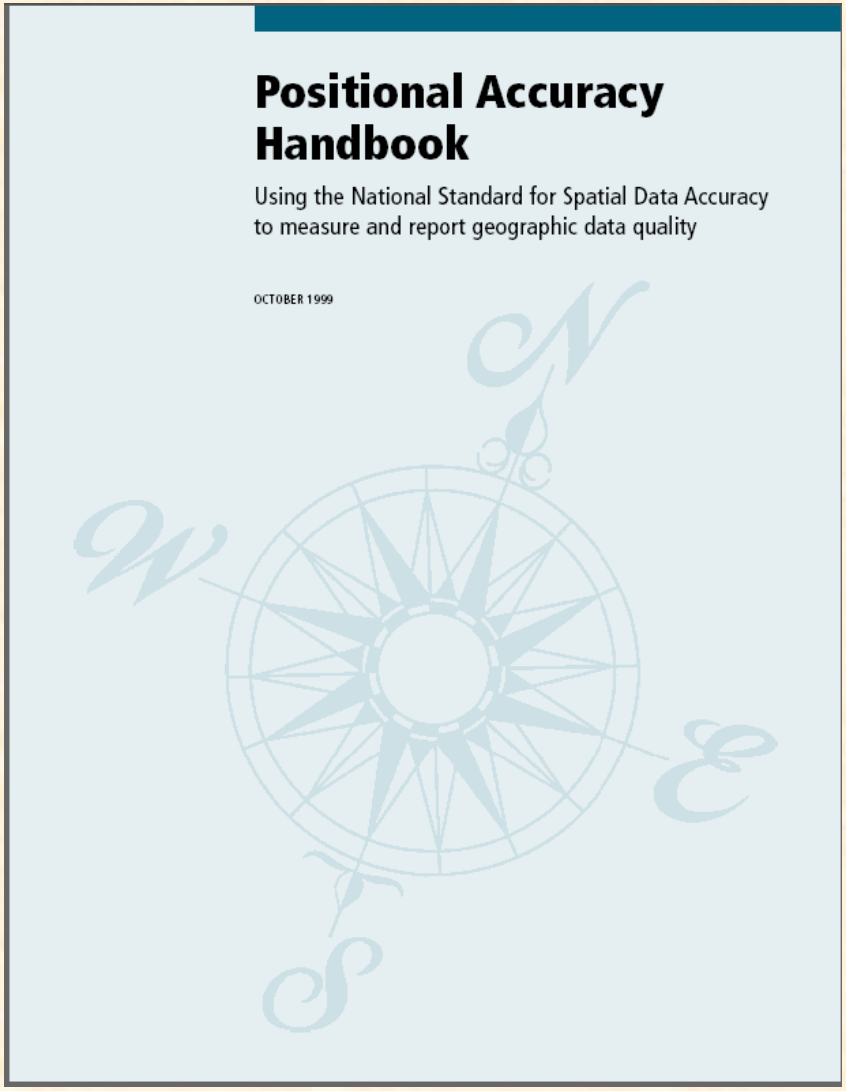
Positional Accuracy

- This presentation only deals with Positional Accuracy.

Need for a Standardized Test

- We must find a way to test the data.
- The test must be:
  - Standardized.
  - Repeatable.
  - Objective.
  - Clearly state the limits of the data.

NSSDA



Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, FL 33410

NSSDA

- National Standard for Spatial Data Accuracy
  - Developed in 1998
  - Issued by the Federal Geographic Data Committee (FGDC)
  - More useful with digital data than NMAS
    - NMAS defines accuracy relative to the map publication scale.

Benefits of the NSSDA

- Identifies a well-defined statistic used to describe test results.
- Describes a method to test spatial data for positional accuracy.
- Provides a common language to report accuracy that makes it easier to evaluate the “fitness for use” of a data set.

Steps for Applying the NSSDA

- Determine which test to use.
- Select a set of test points.
- Select an independent data set.
- Collect measurements.
- Calculate a positional accuracy statistic.
- Prepare accuracy statement.
- Include accuracy statement in metadata.

Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, FL 33410

Determine Which Test to Use

- Horizontal
- Vertical

(Each has a different formula)

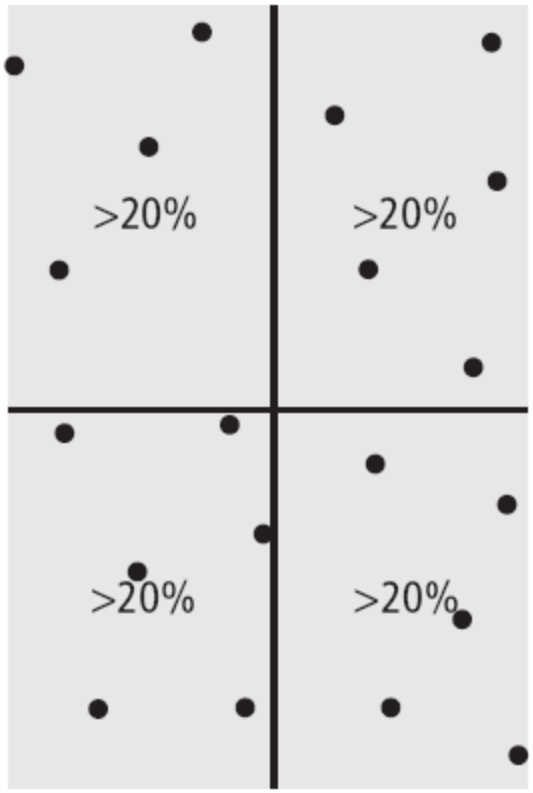
Step 1 - Select Test Points

- Points that will be tested.
- Points must be:
  - Well-defined.
  - Easy to find and measure.
  - Exist in both data sets.
  - Well distributed in data sets.
  - At least 20 points are required to develop a statistically significant accuracy evaluation.

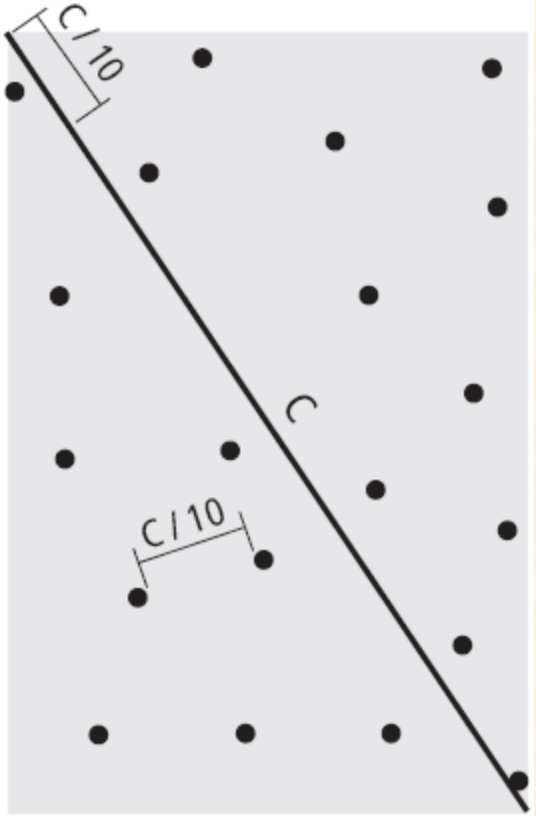
Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, FL 33410

Ideal Distribution of Test Points

Distribution



Spacing



Step 2 - Select Independent Data Set

- Also called the control set.
- The most critical step.
- Data set must be acquired separately from test set.
- Should be 3 times more accurate than the expected accuracy of the test data set.
- If non-existent, data points may be collected and used.

Choosing the Best Data Set

- Since the process compares coordinate values between two data sets, you must ensure that:
  - The data sets are not dependent or derived from one another.
  - The independent data set is accurate.
    - (PLSS/LABINS)
    - Digital Ortho Photography
- Are you testing for absolute or relative accuracy?

Step 3 -Record Measurements

- Collect test point coordinate values from both sets.
- Be aware of errors associated with collection techniques.
  - COGO
  - Digitizing from aerials
  - Building from other data
- Be aware of scale at which you acquire coordinates.
- Snap!

Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, FL 33410

Step 4 -Calculate the Statistic

- To develop the NSSDA, 3 statistics are used:
  - The sum of the set of squared differences in coordinate values
  - The average of the sum
  - The root mean square error
    - (The square root of the average)

Step 5 - Calculate the NSSDA

- Multiply the root mean square error (RMSE) of observations by a value that represents the standard error of the mean at the 95 percent confidence level.
- 1.7308 for horizontal accuracy

Standard Horizontal Accuracy Worksheet

A Point number	B Point description	C x (inde-pendent)	D x (test)	E diff in x	F $(\text{diff in } x)^2$	G y (inde-pendent)	H y (test)	I diff in y	J $(\text{diff in } y)^2$	K $(\text{diff in } x)^2 + (\text{diff in } y)^2$

sum	
average	
$RMSE_r$	
NSSDA	

Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, FL 33410

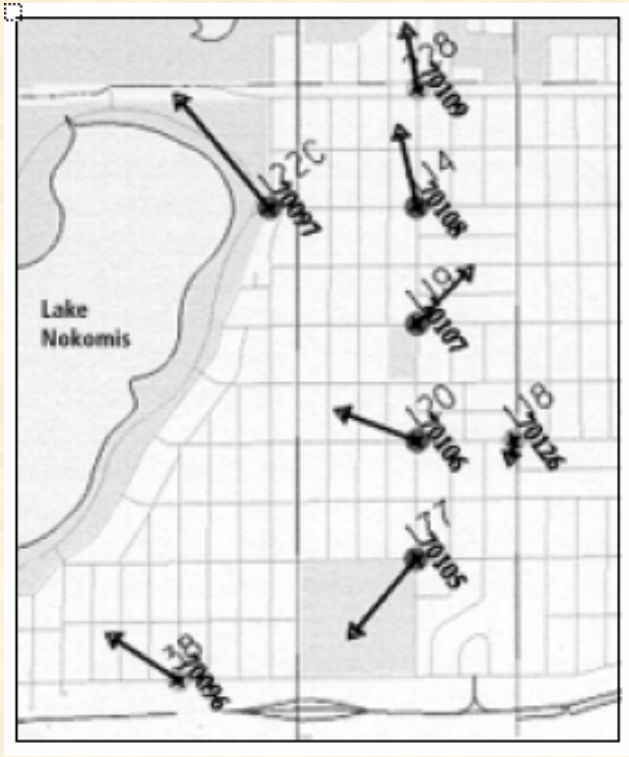
Legend for Worksheet

Column	Title	Content
A	Point number	Designator of test point
B	Point description	Description of test point
C	x (independent)	x coordinate of point from independent data set
D	x (test)	x coordinate of point from test data set
E	diff in x	x (independent) - x (test)
F	(diff in x) ^{2}	Squared difference in x = (x (independent) - x (test)) ^{2}
G	y (independent)	y coordinate of point from independent data set
H	y (test)	y coordinate of point from test data set
I	diff in y	y (independent) - y (test)
J	(diff in y) ^{2}	Squared difference in y = (y (independent) - y (test)) ^{2}
K	(diff in x) ^{2} + (diff in y) ^{2}	Squared difference in x plus squared difference in y = (error radius)^2
	sum	$\sum$ [(diff in x) ^{2} + (diff in y) ^{2}]
	average	sum / number of points
	$RMSE_r$	Root Mean Square Error (radial) = average$^{1/2}$
	NSSDA	National Standard for Spatial Data Accuracy statistic = 1.7308 * $RMSE_r$

Panda Consulting
9089 N. Military Trail, Suite 21 Palm Beach Gardens, FL 33410

What Is RMSE?

- Calculating the average error in location.



- Looks like Georeferencing tool doesn't it?

Sample Completed Worksheet

Point #	Point description	x (inde-pendent)	x (test)	diff ln x	(diff ln x)^2	y (inde-pendent)	y (test)	diff ln y	(diff ln y)^2	(diff ln x)^2 + (diff ln y)^2
10751	r/w & lot line (m&b)	466062.125	466061.709	0.4	0.2	166899.106	166898.974	0.1	0.0	0.2
1100	r/w & lot line (platted)	460363.263	460360.433	2.8	8.0	165103.426	165103.496	-0.1	0.0	8.0
11730	r/w & lot line (m&b)	491133.630	491133.362	0.3	0.1	153041.796	153041.626	0.0	0.0	0.1
1302	r/w & lot line (platted)	462616.265	462616.057	0.2	0.0	166767.766	166767.674	-0.1	0.0	0.1
1397	r/w & lot line (platted)	470569.679	470566.959	0.9	0.8	166328.072	166325.627	0.2	0.1	0.9
1490	r/w & lot line (m&b)	492361.275	492361.352	-0.1	0.0	166191.526	166191.305	0.2	0.0	0.1
2901	r/w & lot line (m&b)	457165.209	457165.039	0.2	0.0	159005.509	159005.616	0.0	0.0	0.0
6160	r/w & lot line (platted)	481796.422	481795.988	0.4	0.2	172592.941	172593.182	-0.2	0.0	0.2
7100	r/w & lot line (platted)	466852.141	466851.230	0.9	0.8	162901.920	162901.132	0.8	0.6	1.5
lot_1_2	r/w & lot line (platted)	451423.044	451422.194	0.8	0.7	173240.565	173240.547	0.3	0.1	0.8
11640	r/w & lot line (platted)	491613.968	491613.949	0.0	0.0	147705.306	147705.645	-0.3	0.1	0.1
3960	r/w & lot line (platted)	453922.111	453922.118	0.0	0.0	153176.492	153176.429	0.1	0.0	0.0
4041	r/w & lot line (platted)	479920.567	479920.492	0.1	0.0	152711.577	152711.656	0.0	0.0	0.0
5120	r/w & lot line (platted)	475454.065	475453.940	0.1	0.0	147133.055	147133.256	-0.2	0.0	0.0
5549	r/w & lot line (platted)	469407.975	469407.927	0.0	0.0	144450.696	144450.912	-0.2	0.0	0.0
6391	r/w & lot line (platted)	463062.352	463062.426	-0.1	0.0	143447.557	143447.761	-0.2	0.0	0.0
6576	r/w & lot line (platted)	463613.337	463613.443	-0.1	0.0	155699.945	155700.107	-0.2	0.0	0.0
8009	r/w & lot line (platted)	472135.343	472135.103	0.2	0.1	153996.576	153996.464	0.1	0.0	0.1
9036	r/w & lot line (platted)	476399.063	476399.053	0.0	0.0	157767.856	157767.940	-0.1	0.0	0.0
9078	r/w & lot line (platted)	475640.112	475639.711	0.4	0.2	146370.597	146370.816	-0.2	0.0	0.2
4766	r/w & lot line (platted)	465173.302	465173.120	0.2	0.0	146308.282	146308.520	-0.3	0.1	0.1
									sum	12.5
									average	0.6
									RMSE	0.5
									NSSDA	1.3

Step 6 - Prepare Statement

- One of two reporting methods can be used:
  - Tested ____ (feet) horizontal accuracy at 95% confidence level.
  - Compiled to meet ____ (feet) horizontal accuracy at 95% confidence level.

Include Accuracy Statement in Metadata

- Sample text:

Digitized features outside areas of high vertical relief: tested 23 feet horizontal accuracy at the 95% confidence level using the NSSDA.

Digitized features within areas of high vertical relief (such as major river valleys): tested 120 feet horizontal accuracy by other testing procedures.

For a complete report of the testing procedures used, contact Washington County Surveyor's Office as noted in Section 6, Distribution Information.

All other features are generated by coordinate geometry and are based on a framework of accurately located PLSS corner positions used with public information of record. Computed positions of parcel boundaries are not based on individual field surveys. Although tests of randomly selected points for comparison may show high accuracy between field and parcel map content, variations between boundary monumentation and legal descriptions of record can and do exist. Caution is necessary when using land boundary data shown. Contact the Washington County Surveyor's Office for more information.

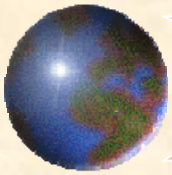
Issues

- Comparative in nature
  - Problem of positional accuracy versus relative accuracy.
  - Data sets can be accurate to one another but positionally inaccurate.
- Only tests single data sets.
- Many data sets are derivative in nature and are dependent upon others for spatial lineage.

- Error is usually not uniform across entire data set.
  - Due to higher concentration of “control points,” suburban and urban areas are often more accurate than rural areas.
  - It is better to perform several tests and state the differences than to conduct a single test and develop NSSDA to entire data set.

Positional Dependency

- The location of any single parcel is dependent upon the features used to locate that parcel.
  - Proven and unproven PLSS corners
  - Other parcels
  - Street centerlines
  - Subdivision boundaries
- In a parcel data set, a single incorrectly located PLSS section corner can impact all parcels within a half-mile radius.
  - The 4 adjoining sections formed from the location of that corner can be impacted.



Data Fabrics

- Often, data is built with implicit dependencies.
- Also called “topology rules”

Therefore…

- It is rare that single data sets can be spatially corrected without impacting many other data sets.
- Often, complete re-engineering projects must be undertaken.

Re-engineering Projects

- When planning re-engineering projects:
  - Recognize different accuracy needs
    - Urban
    - Suburban
    - Rural
  - Ease in acquiring control points.
  - Mapping versus survey control points.
  - Take a long term approach.
    - (Tom Sawyer approach)

Questions?

Frank J. Conkling

frank@pandaconsulting.com

561-691-3277