

PART 2: Tool Development

How to Develop Objective Forecast Tools to Predict Air Quality

Joe Cassmassi

South Coast Air Quality Management District

Overview

- Tool/Model Background
- Developing a Database
- Checking the Database
- Creating a Model
 - Regression
 - Decision Trees
 - Model Evaluation
- Alternative Models
- Daily Operation

Model Background – Why Build or Revise a Model?

- Provides a tool for the forecaster that translates weather to a pollution prediction
- Speeds up the forecast process
- Levels the playing field
 - Algorithms set bounds on the prediction
 - Minimize bias among a group of forecasters
- Addresses the changing air quality trend – empirical relationships are outdated
- Incorporates new sources of meteorological or air quality data

Model Background – Basic Model Structure

- Subjective day-in-advance prediction (phenomenological forecasting using conceptual models)
- Objective same-day prediction with perfect meteorological forecast
- Objective day-in-advance prediction (with/without same-day model)
- Subjective override – fine-tune prediction

Model Background – General Conceptual Model

- Define the mechanics of the pollutant episode by evaluating previous events
- Identify key predictor variables
 - Surface: T, Td, RH, dd, ff
 - Aloft: HHH, T, Td, RH, dd, ff
 - Air quality persistence
- Determine timing of data availability
 - Observations before forecast preparation
 - Prognostic forecast verification periods

Model Background – Conceptual PM_{2.5} Model for L.A. Basin

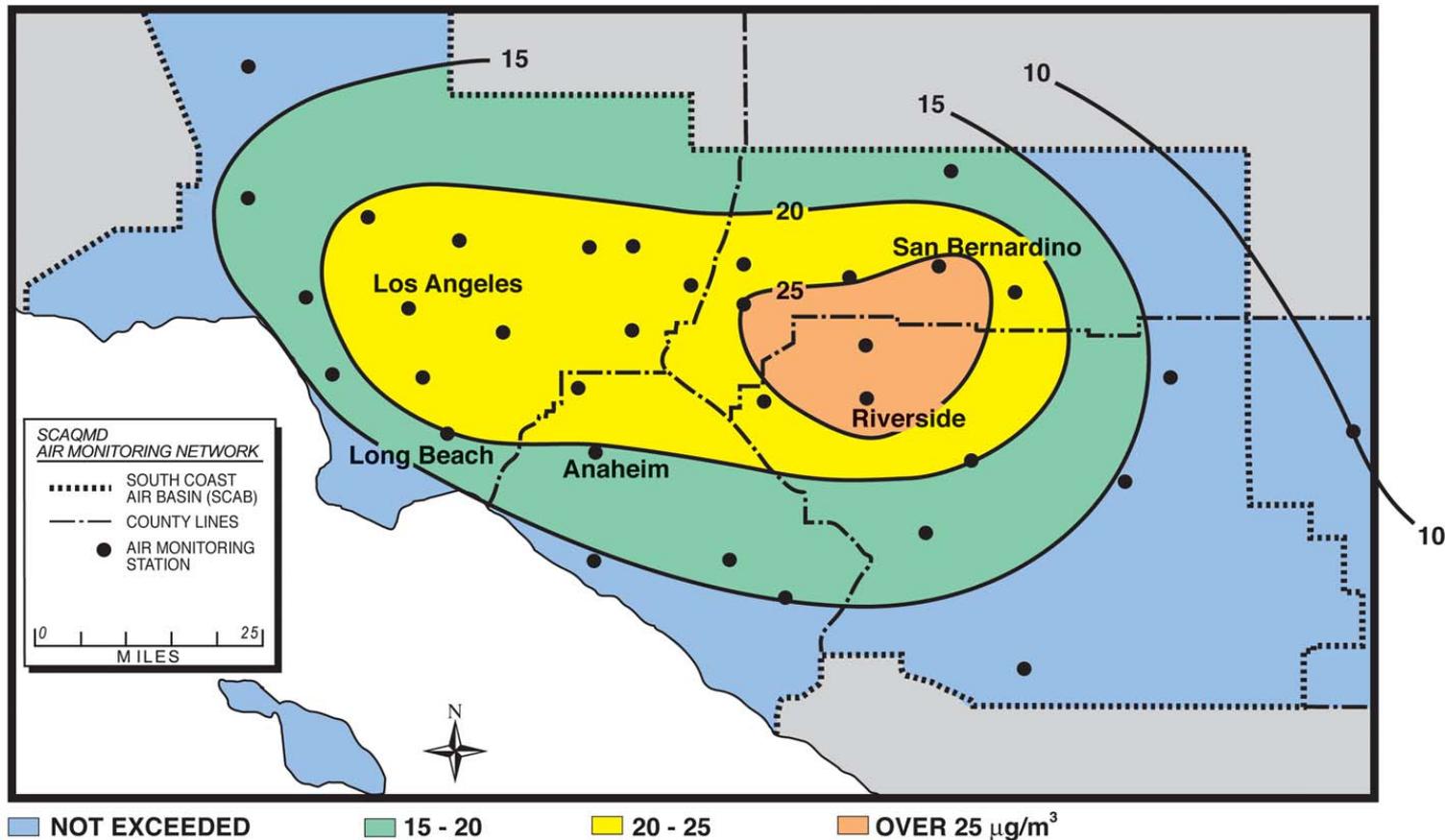
- Gradient of particulates highest in east basin
- PM_{2.5} dominated by secondary PM_{2.5} (nitrate, ammonium and organic carbon [OC]) formed through photochemistry or heterogonous chemistry
- Sulfate, elemental carbon (EC), and crustal contribute less

Conceptual PM_{2.5} Model for L.A. Basin

Spatial Distribution

PM_{2.5} – 2002

Annual Arithmetic Mean, $\mu\text{g}/\text{m}^3$
(Federal Standard = $15 \mu\text{g}/\text{m}^3$)



Conceptual PM_{2.5} Model for L.A. Basin Temporal

- Seasonal
 - All species highest during fall stagnation
 - OC also high mid- and late summer
- Weekend effect
 - Altered VOC/NO_x ratio Saturday and Sunday
 - Lower EC (diesel) on weekend days

Basin PM/Ozone Model – Data Requirements

- Surface: O₃, NO₂, PM₁₀, PM_{2.5}, T, RH, PPP-Gradients
- Aloft: Height (HHH), T, Td, RH (950, 900, 850, 700, 500 mb surfaces)
- Inversion characteristics (top and base HHH and T), stability, thickness
- Prognostic Grid Point FOUS and MOS surface T and RH predictions

Developing a Database

How Much and What Types of Data?

- For statistical analysis, two years minimum
- Data obtained from many sites
- Determined by size and characteristics of forecast domain
- Follow the conceptual model definition
 - Use all available data and let the analysis sort the key predictors
 - Include predictor variables for multiple pollutants (common predictors for O_3 , PM_{10} and $PM_{2.5}$)
 - Include other pollutants (may act as surrogates)

Developing a Database

Locate Sources of Data

- Internally generated air quality
- FSL – upper-air
- Universities (e.g., Wyoming, OSU, PSU, Plymouth State)
- Sister agencies (APCDs, water districts, RAWS, BLM)
- NCDC

FSL RAOB

http://raob.fsl.noaa.gov/temp/raob_soundings25033.tmp - Microsoft Internet Explorer provided

File Edit View Favorites Tools Help Address http://raob.fsl.noaa.gov/temp/raob_soundings25033.tmp Go

254	12	15	JAN	2004		
1	93214	72393	34.75N	120.57W	100	99999
2	70	70	2610	108	99999	3
3		VBG			49	kt
9	10040	100	114	112	320	10
4	10000	132	116	114	335	13
5	9890	224	110	108	99999	99999
5	9880	232	110	107	99999	99999
5	9840	266	110	74	99999	99999
5	9800	300	128	28	99999	99999
6	9795	304	99999	99999	5	12
5	9700	386	144	4	99999	99999
5	9680	403	142	2	99999	99999
6	9446	609	99999	99999	20	19
4	9250	786	120	0	10	15
5	9240	795	120	0	99999	99999
6	9108	914	99999	99999	345	17
5	8970	1041	104	54	99999	99999
6	8779	1219	99999	99999	345	26
5	8670	1322	92	2	99999	99999
5	8580	1408	88	-12	99999	99999
4	8500	1489	82	-18	345	23
6	8154	1828	99999	99999	335	22
5	7930	2055	40	-70	99999	99999
6	7853	2133	99999	99999	330	23
6	7560	2438	99999	99999	325	22
5	7520	2481	-1	-61	99999	99999
5	7400	2609	-11	-49	99999	99999
5	7330	2684	-15	-47	99999	99999

Done Internet

Decoded University of Wyoming RAOB

http://weather.uwyo.edu/cgi-bin/sounding?region=naconf&TYPE=TEXT:LIST&YEAR=2004&MONTH

Address http://weather.uwyo.edu/cgi-bin/sounding?region=naconf&TYPE=TEXT%3ALIS'

72393 VBG Vandenberg Afb Observations at 12Z 15 Jan 2004

PRES hPa	HGHT m	TEMP C	DUPT C	RELH %	MIXR g/kg	DRCT deg	SKNT knot	THTA K	THTE K	THTV K
1004.0	121	11.4	11.2	99	8.38	320	10	284.2	307.5	285.7
1000.0	132	11.6	11.4	99	8.53	335	13	284.8	308.5	286.2
989.0	225	11.0	10.8	99	8.29	351	13	285.1	308.2	286.5
988.0	233	11.0	10.7	98	8.24	353	12	285.1	308.1	286.6
984.0	267	11.0	7.4	78	6.60	358	12	285.5	304.1	286.6
980.0	301	12.8	2.8	51	4.80	4	12	287.6	301.5	288.4
979.5	305	12.9	2.7	50	4.76	5	12	287.7	301.5	288.6
970.0	387	14.4	0.4	38	4.08	9	14	290.1	302.1	290.8
968.0	405	14.2	0.2	38	4.03	10	14	290.0	301.9	290.7
944.6	610	13.0	0.1	41	4.10	20	19	290.9	303.0	291.6
925.0	786	12.0	0.0	44	4.15	10	15	291.6	303.9	292.3
924.0	795	12.0	0.0	44	4.16	8	15	291.7	304.0	292.4
910.9	914	11.2	2.6	55	5.09	345	17	292.1	307.0	293.0
897.0	1043	10.4	5.4	71	6.31	345	21	292.5	310.9	293.6
878.2	1219	9.7	2.1	60	5.12	345	26	293.5	308.6	294.4
867.0	1325	9.2	0.2	53	4.50	345	25	294.1	307.5	294.9
858.0	1412	8.8	-1.2	49	4.10	345	24	294.6	306.9	295.3
850.0	1489	8.2	-1.8	49	3.96	345	23	294.7	306.6	295.4
815.5	1829	5.7	-4.9	46	3.27	335	22	295.6	305.6	296.2
793.0	2058	4.0	-7.0	45	2.86	331	23	296.1	305.0	296.6
785.6	2134	3.3	-6.8	47	2.93	330	23	296.2	305.2	296.7
756.5	2438	0.4	-6.2	61	3.19	325	22	296.2	306.0	296.8

Done Internet

Decoded Surface Observations

http://weather.uwyo.edu/cgi-bin/wyowx.fcgi?TYPE=sflist&DATE=current&HOUR=current&UNITS=

Address: owx.fcgi?TYPE=sflist&DATE=current&HOUR=current&UNITS=A&STATION=LAX

Observations for LOS ANGELES INTL, CA (LAX)

0150Z 15 Jan 2004 to 0050Z 16 Jan 2004

STN	TIME	PMSL	ALTM	TMP	DEW	RH	DIR	SPD	VIS	CLOUDS	Weather	MIN	MAX
	DD/HHMM	hPa	inHg	F	F	%	deg	kt	mile			F	F
LAX	16/0050	1010.7	29.85	58	54	87	270	7	6	FEW250	F		
LAX	15/2350	1010.4	29.84	59	55	87	280	8	6	FEW010 SCT250	F	59	69
LAX	15/2250	1010.4	29.84	60	55	83	270	11	8	FEW200 SCT250			
LAX	15/2150		29.84	66	54	64	270	12	8	FEW200 SCT250			
LAX	15/2050		29.85	68	46	46		5	10	FEW110 FEW200 SCT250			
LAX	15/1950	1011.8	29.88	66	52	61	100	3	10	FEW110 SCT180 BKN220			
LAX	15/1850	1012.8	29.91	65	53	65	120	7	10	FEW110 SCT180 BKN220			
LAX	15/1750	1012.9	29.92	63	53	70	100	5	10	FEW110 SCT180 BKN220		51	63
LAX	15/1650		29.91	59	39	48	100	4	10	FEW080 BKN180 BKN220			
LAX	15/1550		29.89	55	46	72	100	3	10	FEW080 BKN180 BKN220			
LAX	15/1450	1011.8	29.88	57	45	64	0	0	10	SCT080 BKN150			
LAX	15/1350	1011.2	29.87	53	45	74	120	4	10	SCT150			
LAX	15/1250	1011.1	29.86	52	43	71	80	5	10	SCT250			
LAX	15/1150	1011.8	29.88	54	40	59	0	0	10	SCT250		52	66
LAX	15/1050	1012.0	29.89	56	40	55	60	3	10	SCT250			
LAX	15/0950	1012.1	29.90	55	46	72	80	5	10	SCT250			
LAX	15/0850	1011.7	29.88	58	39	49	160	3	10	CLR			
LAX	15/0750	1011.9	29.89	62	41	46		3	10	CLR			
LAX	15/0650	1012.1	29.90	61	44	54	60	4	10	CLR			
LAX	15/0550	1012.2	29.90	59	50	72	70	4	10	FEW250		59	67

AQMD Temperature Data

Station ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0745181980514	38	37	37	38	37	37	38	39	41	42	41	44	44	44	45	45	45	44	43	43	42	42	42	42	42
0745181980515	42	41	41	39	37	38	41	43	47	51	52	53	53	54	54	54	54	51	49	48	48	48	47	47	47
0745181980516	47	47	46	45	46	46	46	46	48	48	49	50	50	51	51	50	49	49	48	46	45	46	46	45	45
0745181980517	45	45	45	43	37	35	38	48	52	55	56	57	58	58	59	59	58	56	54	52	52	51	47	47	47
0745181980518	50	46	44	43	42	43	51	57	59	62	63	63	64	64	64	64	62	59	57	55	56	55	53	53	53
0745181980519	53	53	54	54	54	54	55	57	58	59	60	60	61	61	62	61	61	57	54	52	51	50	50	49	49
0745181980520	49	51	50	50	48	50	51	56	54	54	53	54	55	55	54	54	52	50	47	46	45	44	44	44	44
0745181980521	42	40	38	37	37	37	42	46	48	53	54	54	52	54	54	53	51	50	48	47	46	47	47	46	46
0745181980522	45	43	42	42	41	42	45	47	49	53	56	58	58	57	56	56	53	51	50	49	47	46	46	46	46
0745181980523	46	45	44	43	42	42	46	51	53	56	58	55	51	53	55	54	52	51	50	48	46	48	48	48	48
0745181980524	48	48	45	43	42	43	48	55	57	60	63	66	67	64	61	58	56	53	52	53	54	53	52	51	51
0745181980525	50	50	49	49	48	48	48	49	49	49	50	51	51	50	49	49	50	50	49	49	49	50	49	48	48
0745181980526	48	47	46	46	45	44	44	44	43	43	43	44	43	43	43	45	46	45	43	41	40	40	40	40	40
0745181980527	39	39	38	37	36	36	43	47	46	47	49	51	52	53	54	54	54	52	51	51	52	53	51	49	49
0745181980528	49	48	48	49	49	49	50	51	51	53	53	55	56	57	57	56	55	53	50	48	46	46	45	45	45
0745181980529	45	45	44	45	45	45	46	47	47	46	47	49	50	51	51	51	51	50	50	48	46	49	48	47	47
0745181980530	46	43	42	43	43	40	42	49	54	57	57	58	58	59	59	59	58	54	52	50	51	51	50	50	50
0745181980531	52	52	52	52	51	45	49	56	58	59	60	60	60	61	61	61	61	59	56	56	56	56	56	56	56
0745181980601	56	56	55	51	45	45	53	58	58	60	61	61	62	63	64	64	63	60	57	55	55	55	53	52	52
0745181980602	53	53	52	52	51	52	53	55	56	59	58	58	59	59	59	58	57	54	52	50	49	48	47	45	45
0745181980603	45	44	44	44	44	44	44	44	44	45	45	46	47	47	47	47	47	46	45	44	44	44	44	44	45
0745181980604	45	46	46	46	46	46	45	45	45	45	46	48	49	50	51	52	52	50	49	48	46	46	46	44	44
0745181980605	43	43	42	41	41	42	47	53	60	65	65	65	64	64	63	62	60	58	58	56	56	56	55	54	54
0745181980606	51	51	52	52	52	52	55	56	57	59	59	52	50	51	50	48	47	46	45	45	44	43	42	42	42
0745181980607	41	41	41	41	40	40	40	41	42	42	43	44	44	44	44	43	44	44	45	45	45	45	45	45	45
0745181980608	45	45	45	45	45	45	45	47	48	48	47	48	48	47	49	49	49	49	49	49	48	48	47	47	47
0745181980609	46	45	44	44	44	44	46	49	52	52	53	56	56	53	53	54	54	52	51	50	48	47	46	45	45
0745181980610	45	45	45	45	45	46	51	56	58	62	64	62	55	56	54	52	52	52	51	51	50	50	50	50	50
0745181980611	49	50	50	50	51	50	50	51	50	50	50	51	51	50	50	51	51	50	50	49	49	49	48	48	48
0745181980612	47	46	46	46	46	46	46	46	45	46	46	45	46	46	46	47	46	46	45	46	45	45	46	46	46

FOUS Data

Buckeye Wx - Microsoft Internet Explorer provided by South Coast A.Q.M.D

Address: http://twister.sbs.ohio-state.edu/

Home
Current Wx
Columbus Wx
Weather By State
Upper Air
Tropical Wx
Models
Satellite
Radar
Severe Wx
Ohio Wx
Text Data
Non-U.S. Wx

656
FOUW73 KWNO 231200
OUTPUT FROM NGM 12Z JAN 23 04

TPTTR1R2R3	VVCLI	PSDDFF	HHT1T3T5	TPTTR1R2R3	VVCLI	PSDDFF	HHT1T3T5
SFO//721739	-1414	221101	58080909	FAT//190944	00211	201804	59110904
06000772538	00811	223303	56080906	06000191238	01112	211605	56090703
12000635664	00506	182508	54140904	12000261334	00013	182108	56120602
18000979074	01702	163016	50110801	18000622967	01509	163012	53080501
24005976219	-0502	153016	45100700	24000909758	02802	142805	48070398
30001955023	-0304	163212	44090599	30014947111	00801	142911	44060396
36000723441	-2807	143115	47110400	36000914036	-2104	122914	46070196
42000582268	-1614	183424	46070101	42002933047	-2108	153218	46040098
48000643062	-1715	210120	42030100	48000633063	-3111	173615	44010198
LAX//081614	-1311	190408	55161003	RNO//132145	-0307	211707	56040601
06000102011	02110	200817	53131002	06000131657	-0409	212213	55050400
12000192230	01210	182307	55171002	12000242273	00708	162314	55060299
18000252130	00812	163011	55141002	18000869072	00501	162516	48020096
24000482569	01810	153108	51120801	24020999753	01500	132712	45010093
30000798526	00000	153312	48100600	30006965619	-0302	122913	43019990
42000000000	00000	000000	00000000	36001923047	-0307	112816	42009592
48000932661	-2808	163217	49070401	42000733145	-1211	173216	37969291
				48000744334	-3611	210116	34949291
SLC//413452	-0208	302307	46960097	CDC//181621	-0808	251106	52020297
06000404351	01207	262308	49980198	06000221725	-0209	231705	53010297
12000404145	00607	202112	50990197	12000232828	00407	172408	55060196
18000243659	01607	142315	50000096	18000262645	01108	132516	55060095
24000182561	00807	122419	47990095	24000233471	-1207	122716	50030093

Severe Wx Page, incl. Severe Wx Statements

Internet

AQMD Hi-Vol PM₁₀ Data

ID	Date	Hourly Value	Average Value	Maximum Value	Station	Value
33144	2 JAN 02	61	6.5	19.2	SA	0.0
33144	8 JAN 02	65	1.3	14.0	SA	0.1
33144	14 JAN 02	65	2.8	15.1	SA	0.0
33144	20 JAN 02	55	1.3	8.8	SA	0.0
33144	26 JAN 02	50	1.4	8.8	SA	0.0
33144	5 JAN 02	51	0.6	4.3	SA	0.1
33144	11 JAN 02	33	0.6	1.1	SA	0.1
33144	17 JAN 02	55	2.3	8.7	SA	0.0
33144	23 JAN 02	32	0.6	0.7	SA	0.1
33144	29 JAN 02	9	0.5	0.8	SA	0.0
33144	1 FEB 02	41	0.5	1.3	SA	0.1
33144	7 FEB 02	88	2.8	18.9	SA	0.0
33144	13 FEB 02	74	2.3	12.1	SA	0.0
33144	19 FEB 02	53	1.8	12.2	SA	0.2
33144	25 FEB 02	80	2.7	13.6	SA	0.1
33144	4 FEB 02	28	0.6	0.6	SA	0.0
33144	10 FEB 02	30	0.4	0.3	SA	0.0
33144	16 FEB 02	53	5.6	16.6	SA	0.0
33144	22 FEB 02	43	0.8	1.1	SA	0.2
33144	28 FEB 02	69	4.2	14.2	SA	0.0
33144	3 MAR 02	27	0.9	0.3	SA	0.0
33144	9 MAR 02	37	1.1	4.3	SA	0.2
33144	15 MAR 02	33	1.0	2.9	SA	1.1

AQMD Hi-Vol PM_{2.5} Data

Microsoft Excel - PM25 Data 0202 to 0802

File Edit View Insert Format Tools Data Window Help

MS Sans Serif 10 B I U

F16 fx

	A	B	C	D	E	F	G	H	I	J
1	Station	Filter #	Date	PM 25						
2	AnaheimA	11887	02/01/02	14.9						
3	AnaheimA	11888	02/02/02	31.7						
4	AnaheimA	11889	02/03/02	28.8						
5	AnaheimA	11920	02/04/02	15.9						
6	AnaheimA	11921	02/05/02	-9.0						
7	AnaheimA	11972	02/06/02	32.6						
8	AnaheimA	11973	02/07/02	49.9						
9	AnaheimA	11974	02/08/02	42.9						
10	AnaheimA	12009	02/09/02	13.2						
11	AnaheimA	12010	02/10/02	6.6						
12	AnaheimA	12011	02/11/02	8.5						
13	AnaheimA	12037	02/12/02	24.1						
14	AnaheimA	12038	02/13/02	41.9						
15	AnaheimA	12039	02/14/02	27.5						
16	AnaheimA	12040	02/15/02	42.8						
17	AnaheimA	12093	02/16/02	28.1						
18	AnaheimA	12094	02/17/02	6.8						
19	AnaheimA	12095	02/18/02	9.8						
20	AnaheimA	12121	02/19/02	15.7						
21	AnaheimA	12122	02/20/02	36.7						
22	AnaheimA	12123	02/21/02	13.2						
23	AnaheimA	12186	02/22/02	6.8						
24	AnaheimA	12187	02/23/02	12.9						

SampleSummary_Query

Ready NUM

Developing a Database Common Linear Format

- Most statistical software require a linear format
 - one line of data for each day (see Session 1B Part 1)
- The data must be structured (lined up) based on time considerations
 - Air quality data for preceding, current, and next day (d-1, d0, d+1)
 - Meteorological observational data for time periods prior to expected forecast preparation/issuance time
 - Prognostic data (or surrogate prognostic data) for the next day

Developing a Database

Common Predictor/Stratifying Variables

- Day-of-week indicator
(numerical 1,2...7 or binary 0,1)
- Julian day of year (1-365)
- Number of hours in day
(function of Julian day and latitude)
- Holiday marker (binary 0,1)

Developing a Database

Translating and Reformatting the Data

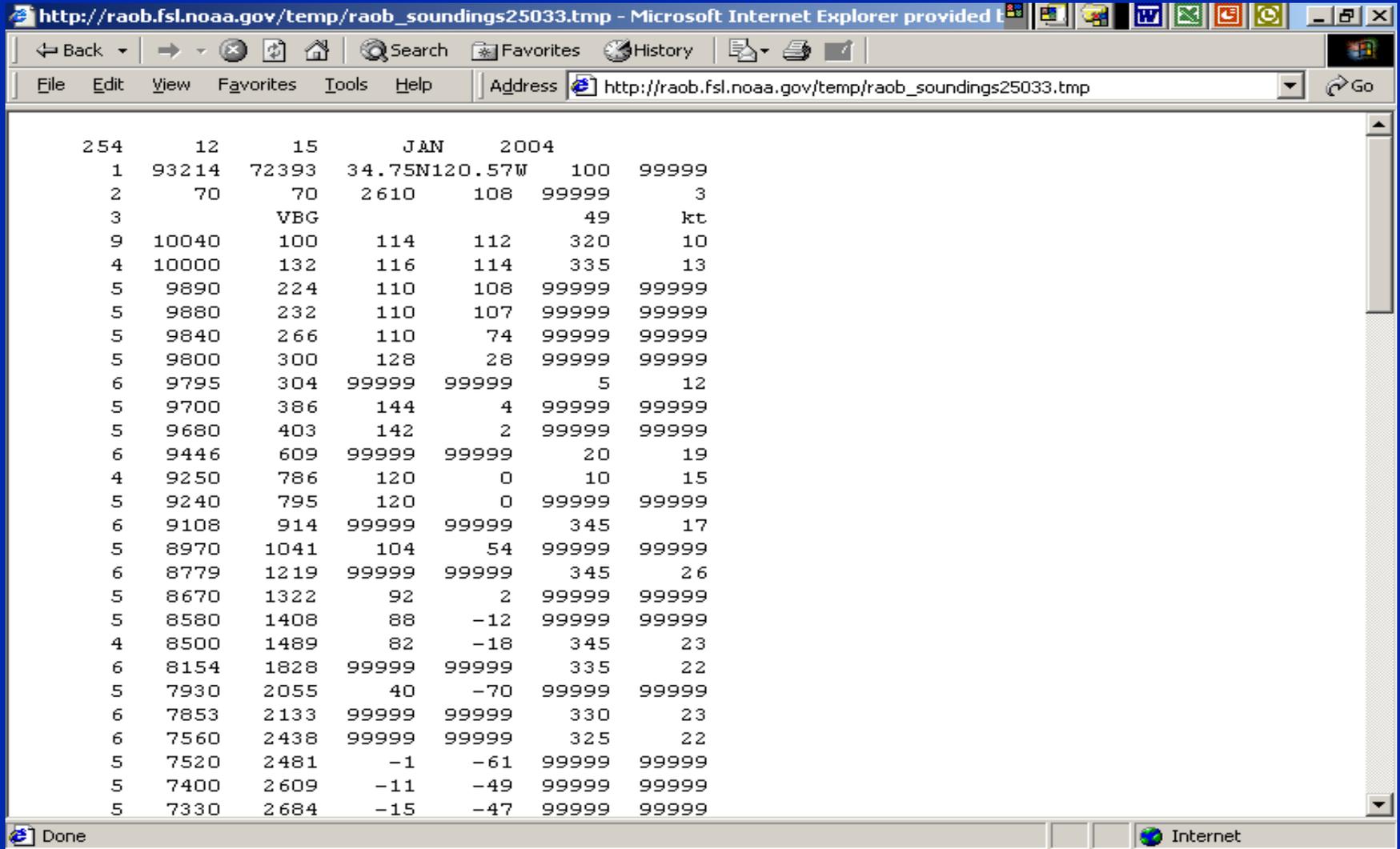
- Only critical rule is to line up the data chronologically (remember one day's data on one record)
- Write simple programs to extract and massage data (Fortran, Basic, C++,...)
- Spreadsheets and databases are good platforms for merging individual or groups of products and aligning dates
- Use as many tools as necessary – use programs and spreadsheets



Developing a Database – Example

Extracting Upper-Air Data From an FSL-formatted RAOB Archive

Base FSL Format



The screenshot shows a Microsoft Internet Explorer browser window displaying a Base FSL format file. The address bar shows the URL: http://raob.fsl.noaa.gov/temp/raob_soundings25033.tmp. The browser's menu bar includes File, Edit, View, Favorites, Tools, and Help. The address bar also contains a search icon, a home icon, and a go button. The main content area displays the following text:

```
254      12      15      JAN      2004
  1  93214  72393  34.75N120.57W  100  99999
  2      70      70      2610      108  99999      3
  3              VBG              49      kt
  9  10040      100      114      112      320      10
  4  10000      132      116      114      335      13
  5  9890      224      110      108  99999  99999
  5  9880      232      110      107  99999  99999
  5  9840      266      110      74   99999  99999
  5  9800      300      128      28   99999  99999
  6  9795      304  99999  99999      5      12
  5  9700      386      144      4   99999  99999
  5  9680      403      142      2   99999  99999
  6  9446      609  99999  99999      20      19
  4  9250      786      120      0      10      15
  5  9240      795      120      0   99999  99999
  6  9108      914  99999  99999      345      17
  5  8970     1041      104      54  99999  99999
  6  8779     1219  99999  99999      345      26
  5  8670     1322      92      2   99999  99999
  5  8580     1408      88     -12  99999  99999
  4  8500     1489      82     -18   345      23
  6  8154     1828  99999  99999      335      22
  5  7930     2055      40     -70  99999  99999
  6  7853     2133  99999  99999      330      23
  6  7560     2438  99999  99999      325      22
  5  7520     2481      -1     -61  99999  99999
  5  7400     2609      -11    -49  99999  99999
  5  7330     2684     -15    -47  99999  99999
```

The status bar at the bottom of the browser window shows "Done" on the left and "Internet" on the right.

Strip Header, Grab Time/Date, Eliminate Unneeded Data

Microsoft Excel - NKX00ZA

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U

K35

	A	B	C	D	E	F	G	H	I	J	K	L
29	0	05-Jan-00	3590	8194	-341	-471	99999	99999				
30	0	05-Jan-00	3269	8839	99999	99999	320	26				
31	0	05-Jan-00	3127	9144	99999	99999	325	26				
32	0	05-Jan-00	3000	9430	-441	-551	325	26				
33	0	05-Jan-00	2600	10363	99999	99999	330	33				
34	0	05-Jan-00	2500	10620	-543	-643	330	32				
35	0	05-Jan-00	2481	10668	99999	99999	330	32				
36	12	05-Jan-00	10020	134	68	-2	80	7				
37	12	05-Jan-00	10000	150	90	0	80	7				
38	12	05-Jan-00	9900	233	124	54	99999	99999				
39	12	05-Jan-00	9816	304	99999	99999	55	4				
40	12	05-Jan-00	9790	326	130	0	99999	99999				
41	12	05-Jan-00	9466	609	99999	99999	330	3				
42	12	05-Jan-00	9460	614	160	-30	99999	99999				
43	12	05-Jan-00	9250	808	146	-34	335	4				
44	12	05-Jan-00	9134	914	99999	99999	345	5				
45	12	05-Jan-00	8808	1219	99999	99999	355	6				
46	12	05-Jan-00	8500	1517	112	-68	330	10				
47	12	05-Jan-00	8220	1795	102	-98	99999	99999				
48	12	05-Jan-00	8187	1828	99999	99999	305	15				
49	12	05-Jan-00	7890	2133	80	-20	99999	99999				
50	12	05-Jan-00	7890	2133	99999	99999	280	17				
51	12	05-Jan-00	7601	2438	99999	99999	290	18				
52	12	05-Jan-00	7510	2537	66	-174	99999	99999				

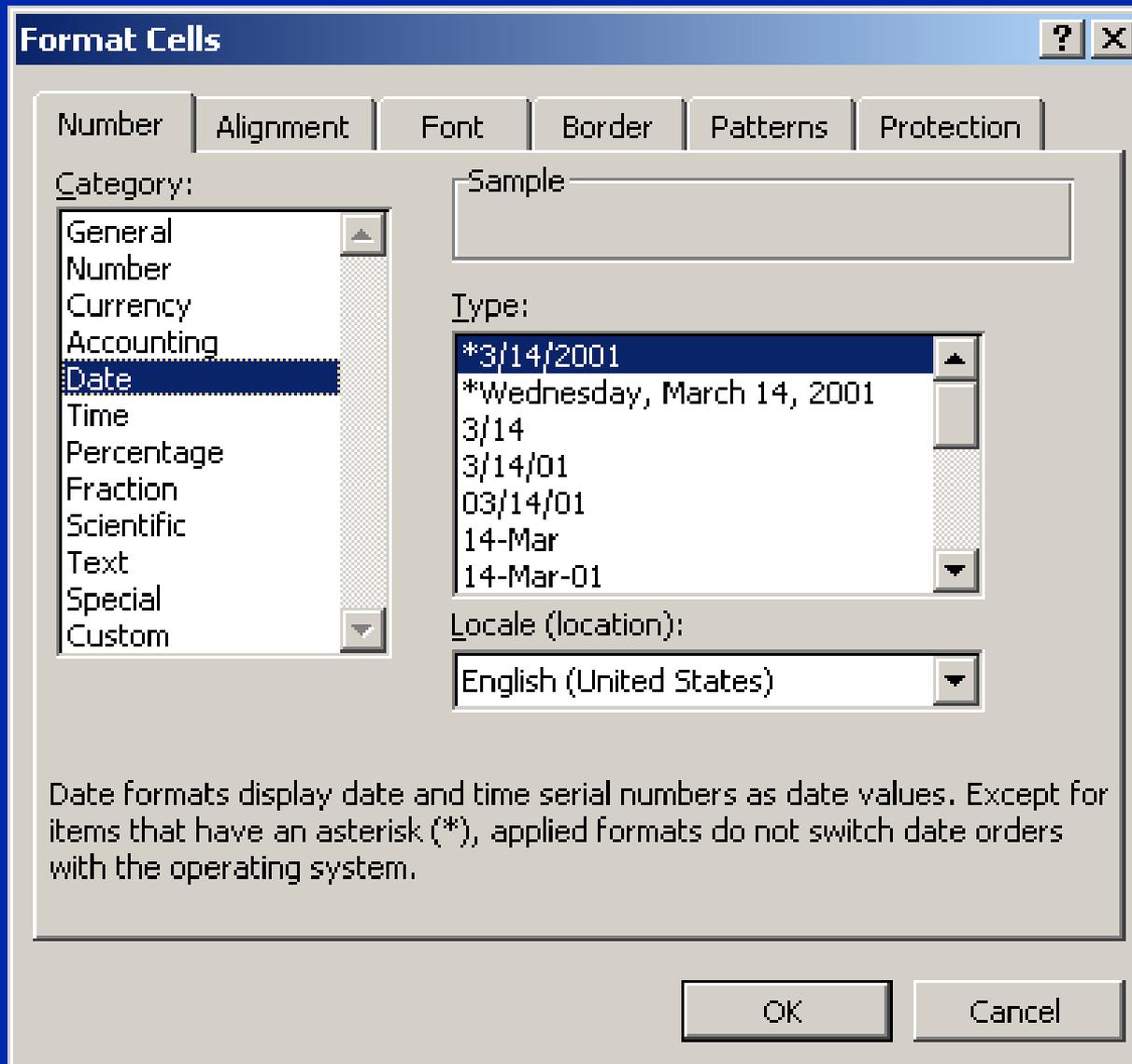
Ready NUM

Convert Date to Numeric

The screenshot shows Microsoft Excel with a spreadsheet containing data from rows 20 to 43. The columns are labeled A through L. The data includes numerical values and dates in the format MM-DD-YY. Cell K35 is selected, and a numeric input box is visible next to it, indicating the user is entering a numeric value for a date.

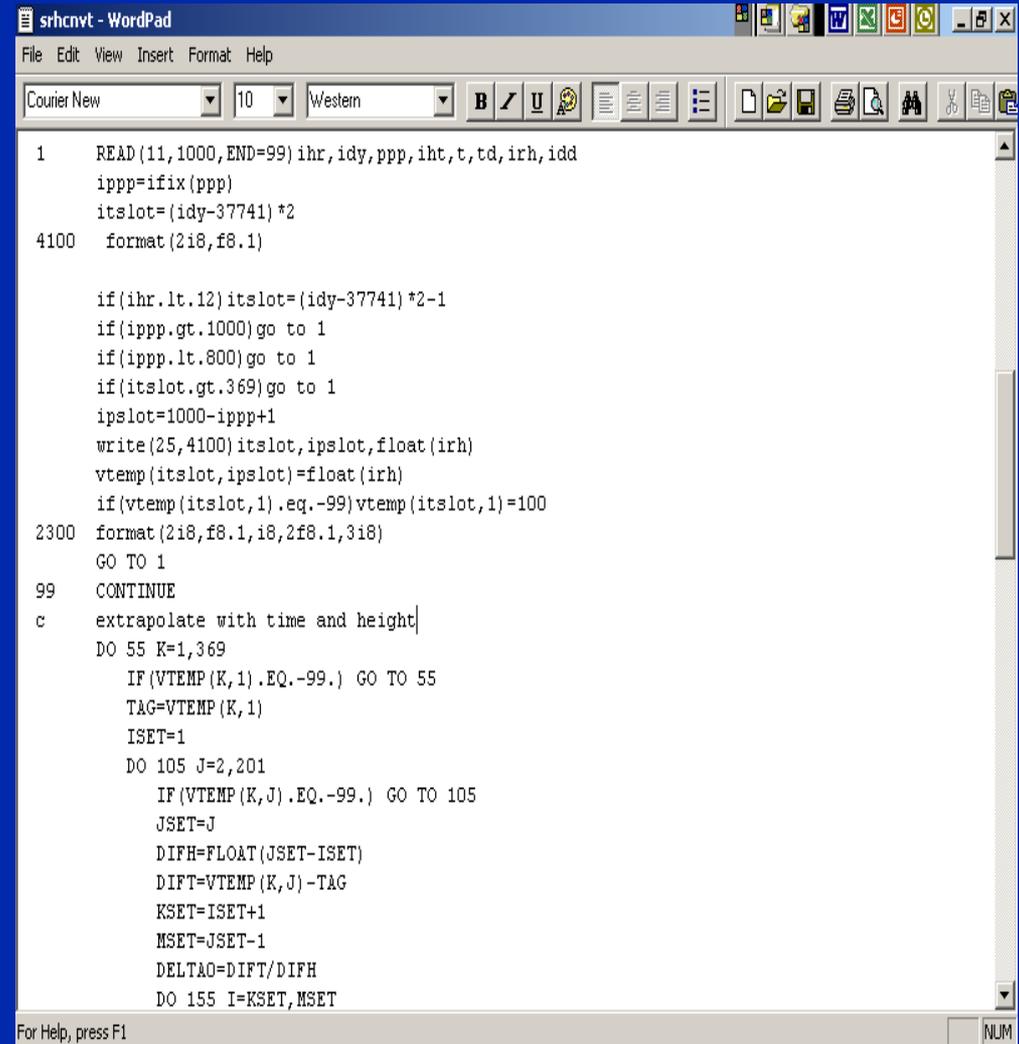
	A	B	C	D	E	F	G	H	I	J	K	L
20	0	36530	5510	5047	-109	-249	99999	99999				
21	0	36530	5220	5459	-141	-211	99999	99999				
22	0	36530	5000	5790	-159	-239	320	23				
23	0	36530	4798	6096	99999	99999	320	24				
24	0	36530	4430	6686	-235	-305	99999	99999				
25	0	36530	4418	6705	99999	99999	315	19				
26	0	36530	4064	7315	99999	99999	285	20				
27	0	36530	4000	7430	-283	-383	285	20				
28	0	36530	3894	7620	99999	99999	290	20				
29	0	05-Jan-00	3590	8194	-341	-471	99999	99999				
30	0	05-Jan-00	3269	8839	99999	99999	320	26				
31	0	05-Jan-00	3127	9144	99999	99999	325	26				
32	0	05-Jan-00	3000	9430	-441	-551	325	26				
33	0	05-Jan-00	2600	10363	99999	99999	330	33				
34	0	05-Jan-00	2500	10620	-543	-643	330	32				
35	0	05-Jan-00	2481	10668	99999	99999	330	32				
36	12	05-Jan-00	10020	134	68	-2	80	7				
37	12	05-Jan-00	10000	150	90	0	80	7				
38	12	05-Jan-00	9900	233	124	54	99999	99999				
39	12	05-Jan-00	9816	304	99999	99999	55	4				
40	12	05-Jan-00	9790	326	130	0	99999	99999				
41	12	05-Jan-00	9466	609	99999	99999	330	3				
42	12	05-Jan-00	9460	614	160	-30	99999	99999				
43	12	05-Jan-00	9250	808	146	-34	335	4				

Excel Date Formatting



Extracting Data Using Numeric Date, Time, and Pressure Tags

- Extrapolate data
 - > vertically (surface to 800 mb)
 - > with time (00Z and 12Z soundings)
- Interpolate throughout year for missing data



```
srhcnvt - WordPad
File Edit View Insert Format Help
Courier New 10 Western B / U
1 READ(11,1000,END=99) ihr, idy, ppp, iht, t, td, irh, idd
  ipp=ifix(ppp)
  itslot=(idy-37741)*2
4100 format(2i8,f8.1)

  if(ihr.lt.12) itslot=(idy-37741)*2-1
  if(ipp.gt.1000) go to 1
  if(ipp.lt.800) go to 1
  if(itslot.gt.369) go to 1
  ipslot=1000-ipp+1
  write(25,4100) itslot, ipslot, float(irh)
  vtemp(itslot, ipslot)=float(irh)
  if(vtemp(itslot,1).eq.-99) vtemp(itslot,1)=100
2300 format(2i8,f8.1,i8,2f8.1,3i8)
  GO TO 1
99 CONTINUE
c extrapolate with time and height
DO 55 K=1,369
  IF (VTEMP(K,1).EQ.-99.) GO TO 55
  TAG=VTEMP(K,1)
  ISET=1
  DO 105 J=2,201
    IF (VTEMP(K,J).EQ.-99.) GO TO 105
    JSET=J
    DIFH=FLOAT(JSET-ISET)
    DIFT=VTEMP(K,J)-TAG
    KSET=ISET+1
    MSET=JSET-1
    DELTAO=DIFT/DIFH
  DO 155 I=KSET,MSET
```


Align Air Quality Data

Station ID	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col 11	Col 12	Col 13	Col 14	Col 15	Col 16	Col 17	Col 18	Col 19
36161	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36162	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36163	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36164	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36165	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36166	63	51	-99	105	75	26	87	-99	23	75	25	71	58	33	97	76	64	62
36167	-99	-99	27	-99	-99	-99	-99	67	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36168	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36169	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36170	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36171	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36172	39	30	27	94	68	39	54	-99	39	65	39	76	73	50	73	51	52	36
36173	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36174	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36175	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36176	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36177	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36178	53	49	38	118	91	47	85	-99	47	92	18	100	93	78	106	91	69	37
36179	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36180	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36181	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36182	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36183	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36184	23	20	33	43	30	10	35	-99	11	35	7	43	37	30	30	26	20	14
36185	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36186	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36187	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36188	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
36189	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99

Missing Data

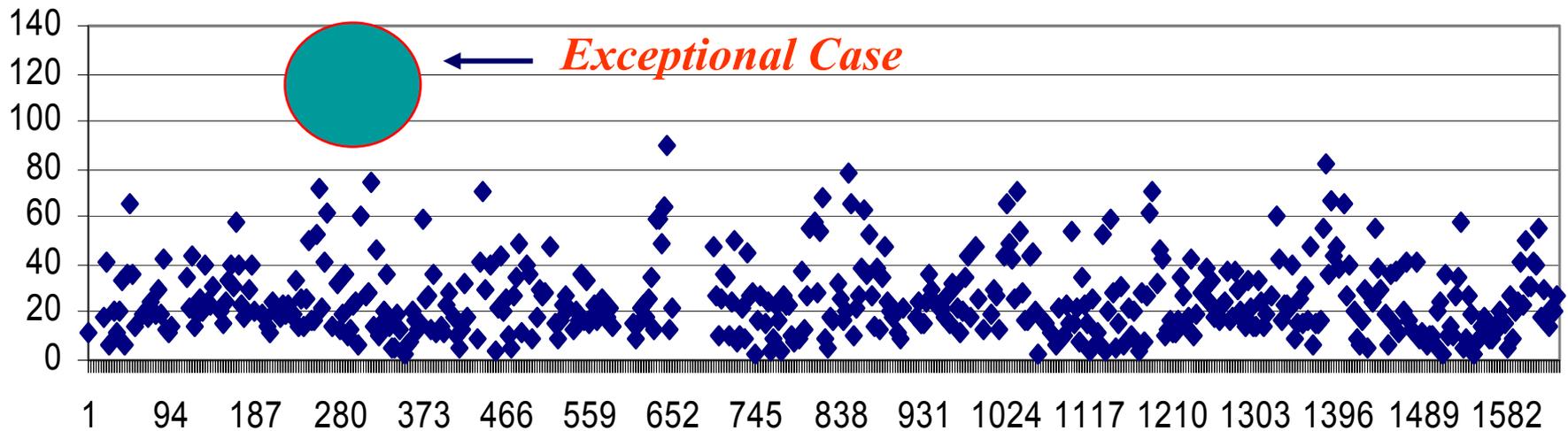
- Missing data must be marked in the data set
- If possible, fill in spurious gaps in the data through extrapolation
- Minimize the use of variables that have large or routine gaps in the data
- Statistical software typically eliminates a record from the analysis if it includes a variable with missing data

Checking the Database

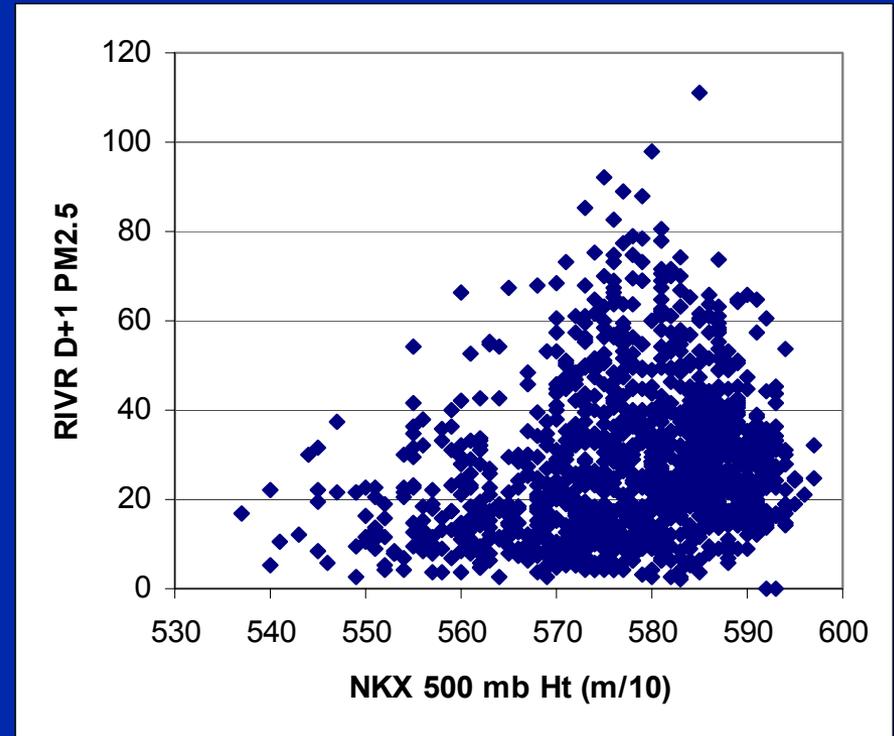
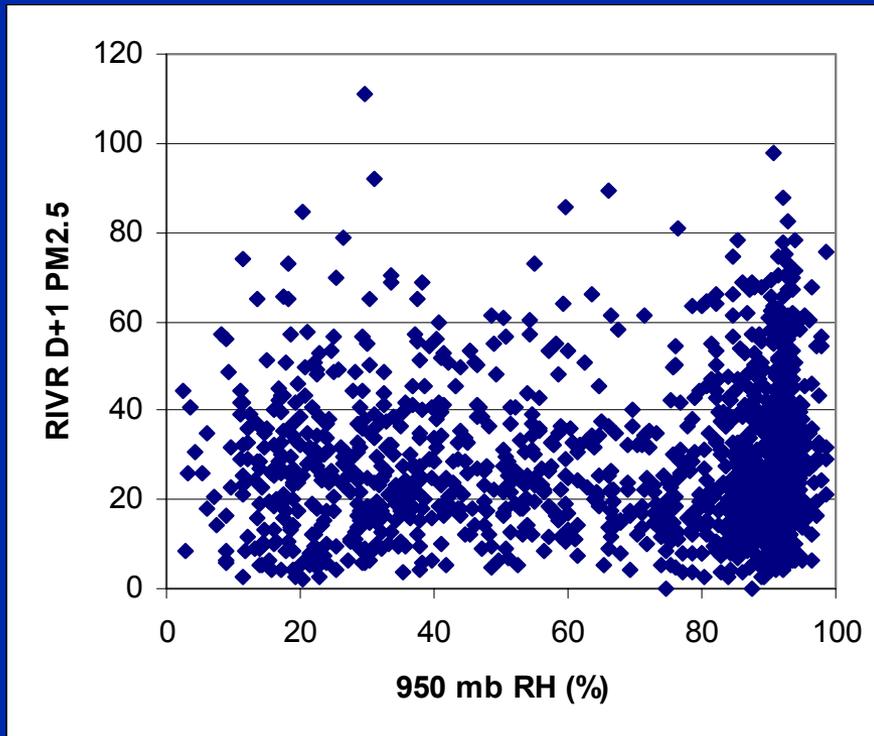
- Use a spreadsheet or canned programs to assess the data
- Basic statistics
 - Maximum, minimum, and range
 - Average
 - Correlation matrices
- Graphical presentations to identify unique data and outliers
 - Time series plots
 - Bivariate data plots

Checking the Database Data Analysis

- Examine trend
- Scan data for outliers
- Note unusual or exceptional cases



Checking the Database Analyze Scatter Plots



Creating a Model

- Overview
- Stepwise Multivariate Regression
- Decision Trees
- Model Evaluation

Overview – Selecting Statistical Software

- Availability, platform, and cost
- What statistical algorithms are provided?
- Data transformation
 - Fixed format (flat ASCII)
 - Spreadsheet or variable format (.xls)
 - Translation software (DBMS Copy)
- Windows-based?
- Output formats
- Ease of use

Overview – Common Statistical Packages

- Excel (Basics)
- SYSTAT
- BMDP
- SASS
- SPSS
- CART
- Neural Networks

Stepwise Multivariate Regression (1 of 2)

- Data Considerations
 - The number of cases is more important than the number of predictors
 - Two years of data at minimum
 - Select the appropriate predictor variables
 - Missing data limits number of cases
 - Sort data into dependent and independent data sets for validation

Stepwise Multivariate Regression (2 of 2)

- Algorithm Fitting Criteria
 - Software sets a default criteria for developing empirical model (probabilistic, f-level to enter...)
 - Lowering the default criteria to allow variables into the algorithm development will result in over-fitting the data
 - Small data samples (only a few records) can result in over-fitting as well

Regression Model Evaluation

- Evaluate R , R^2 , and error
- Considerations
 - What variables are included in the algorithm?
 - Does the algorithm make physical sense?
 - Variables selected consistent with conceptual model?
 - Directional sign of coefficients consistent?
- Assess variable contribution to prediction
 - Value of intercept
 - Weight of coefficients

Stepwise Multivariate Regression – Excel Database (1 of 2)

Microsoft Excel - pmmeta2

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial 10 B I U

BE12 39

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Year	Month	Day	DWKO	DWK1	PO	P1	HOURS	H5VBG	H5SAN	H5DRA	H7VBG	H7SAN	H7DRA	H8VBG	H8SAN	H8DRA
2	98	12	31	4	5	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
3	99	1	1	5	6	3	3	9.72231	579	578	572	315	313	308	154	151	150
4	99	1	2	6	7	4	4	9.72823	584	581	576	317	313	313	155	153	150
5	99	1	3	7	1	3	3	9.73485	580	576	569	315	311	311	153	152	150
6	99	1	4	1	2	4	4	9.74215	583	581	579	319	317	316	158	156	150
7	99	1	5	2	3	3	3	9.75014	584	586	580	320	318	317	158	157	150
8	99	1	6	3	4	3	3	9.75882	582	581	579	319	317	316	157	156	150
9	99	1	7	4	5	3	3	9.76818	575	576	571	313	313	310	153	152	150
10	99	1	8	5	6	4	4	9.77822	576	570	563	311	309	307	152	149	150
11	99	1	9	6	7	4	4	9.78894	582	580	577	316	314	315	156	153	150
12	99	1	10	7	1	4	4	9.80033	582	581	580	317	315	315	155	154	150
13	99	1	11	1	2	3	3	9.81239	578	578	577	313	312	312	153	152	150
14	99	1	12	2	3	3	3	9.82512	574	571	566	311	309	305	152	149	140
15	99	1	13	3	4	3	3	9.83885	577	577	571	313	311	309	153	151	150
16	99	1	14	4	5	3	3	9.85255	586	577	580	317	314	316	158	157	150
17	99	1	15	5	6	3	3	9.86725	585	583	580	317	316	313	156	152	150
18	99	1	16	6	7	3	3	9.8826	573	575	566	310	310	304	152	150	140
19	99	1	17	7	1	3	2	9.8986	577	576	568	313	310	307	153	151	140
20	99	1	18	1	2	3	3	9.91524	576	578	573	313	314	310	154	153	150
21	99	1	19	2	3	3	3	9.93251	570	575	569	309	312	307	152	152	140
22	99	1	20	3	4	3	3	9.95041	566	570	562	308	309	303	151	152	140
23	99	1	21	4	5	5	3	9.96893	572	574	555	310	309	299	152	152	140
24	99	1	22	5	6	3	2	9.98808	579	577	567	314	312	307	155	152	150

pmmeta2

Edit NUM

Stepwise Multivariate Regression – Excel Database (2 of 2)

Microsoft Excel - pmmeta2

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U \$ % , +.00 +.00

39

	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	BASEHT	TOPHT	RH995	RH950	RH900	RH800	ONTA250	RIVR250	RIVS250	RIVR251	RIVS251	RIVR100	RIVRTE	MIRATO
2	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
3	800	3250	86.7	96.5	32.6	18								49.7
4	100	2700	50.9	21.8	19.5	19.8	21.9	6.2						38.2
5	100	1600	46.9	22.6	21.9	21		12.3		6.2				33
6	100	700	27.6	23.8	23	21.4		29.8		12.3				51.1
7	100	1940	21.2	18.4	18.7	17.2	37.3	37.2	36.1	29.8				68.8
8	100	2720	28.6	16.6	15.5	16.2		57.3		37.2	36.1	105		75.8
9	100	2650	68.2	28.9	24.6	24.9				57.3				70.3
10	2000	2700	89.1	94.7	59.6	24.1	39.7	13.8	9.9					69.8
11	100	700	21.1	21.1	21.5	19.1		22		13.8	9.9			61.2
12	100	2440	24.8	20.6	20.4	19.9		28.2		22				45.1
13	1300	2600	50.9	26.4	29.3	28.6	47.5	49.5	49.5	28.2				56
14	1100	1800	91	33.2	33.4	30		78.7		49.5	49.5	94		60.3
15	1500	3000	98.5	80.5	30.5	28.1		20.9		78.7				71.9
16	700	1400	99.2	24.9	24.5	26.3	35.4	31.1	31.1	20.9				53.5
17	100	1300	28.5	25.3	25.6	24.8		56.7		31.1	31.1			74.9
18	500	2000	64.2	30.9	36.9	47.3		70		56.7				67.2
19	2980	6430	84.2	98	46.4	25.2	85.9	92	89.9	70				42.4
20	2500	4500	96.2	74.4	55.6	7		56.3		92	89.9	118		58.1
21	2900	3600	89.4	96.5	22.8	2.5		25.6		56.3				25.1
22	5900	6100	96.6	96.1	95.8	95.7	58.6	17.6		25.6				5.9
23	9998	9999	96.6	96.2	95.6	95.1		26.8		17.6				20.1
24	100	2700	92.7	51.3	35.1	29.5		34.7		26.8				60.7

NUM

Stepwise Multivariate Regression – Data Imported into SYSTAT

Untitled - SYSTAT Data

File Edit View Data Graph Statistics Help

Row: 1, Variable: YEAR -99

	YEAR	MONTH	DAY	DWK0	DWK1	P0	P1	HOURS	H5VBG
1	-99.000	12.000	31.000	4.000	5.000	-99.000	-99.000	-99.000	-99.0
* 2	99.000	1.000	1.000	5.000	6.000	3.000	3.000	9.722	579.0
* 3	99.000	1.000	2.000	6.000	7.000	4.000	4.000	9.728	584.0
* 4	99.000	1.000	3.000	7.000	1.000	3.000	3.000	9.735	580.0
* 5	99.000	1.000	4.000	1.000	2.000	4.000	4.000	9.742	583.0
* 6	99.000	1.000	5.000	2.000	3.000	3.000	3.000	9.750	584.0
* 7	99.000	1.000	6.000	3.000	4.000	3.000	3.000	9.759	582.0
* 8	99.000	1.000	7.000	4.000	5.000	3.000	3.000	9.768	575.0
* 9	99.000	1.000	8.000	5.000	6.000	4.000	4.000	9.778	576.0
* 10	99.000	1.000	9.000	6.000	7.000	4.000	4.000	9.789	582.0
* 11	99.000	1.000	10.000	7.000	1.000	4.000	4.000	9.800	582.0
* 12	99.000	1.000	11.000	1.000	2.000	3.000	3.000	9.812	578.0
* 13	99.000	1.000	12.000	2.000	3.000	3.000	3.000	9.825	574.0
* 14	99.000	1.000	13.000	3.000	4.000	3.000	3.000	9.839	577.0
* 15	99.000	1.000	14.000	4.000	5.000	3.000	3.000	9.853	586.0
* 16	99.000	1.000	15.000	5.000	6.000	3.000	3.000	9.867	585.0
* 17	99.000	1.000	16.000	6.000	7.000	3.000	3.000	9.883	573.0
* 18	99.000	1.000	17.000	7.000	1.000	3.000	2.000	9.899	577.0
* 19	99.000	1.000	18.000	1.000	2.000	3.000	3.000	9.915	576.0
* 20	99.000	1.000	19.000	2.000	3.000	3.000	3.000	9.933	570.0
* 21	99.000	1.000	20.000	3.000	4.000	3.000	3.000	9.950	566.0

Ready

SYSTAT – Variable Sorting

Select Cases [X]

Variable(s):
Case
YEAR
MONTH
DAY
DWK0
DWK1
P0

Function Type:
Mathematical

Functions:
SQR
LOG
L10
EXP

Variable(s):
Case
YEAR
MONTH
DAY
DWK0
DWK1
P0

Add Add Add

Variable Variable or Function

Select YEAR > -1

AND =

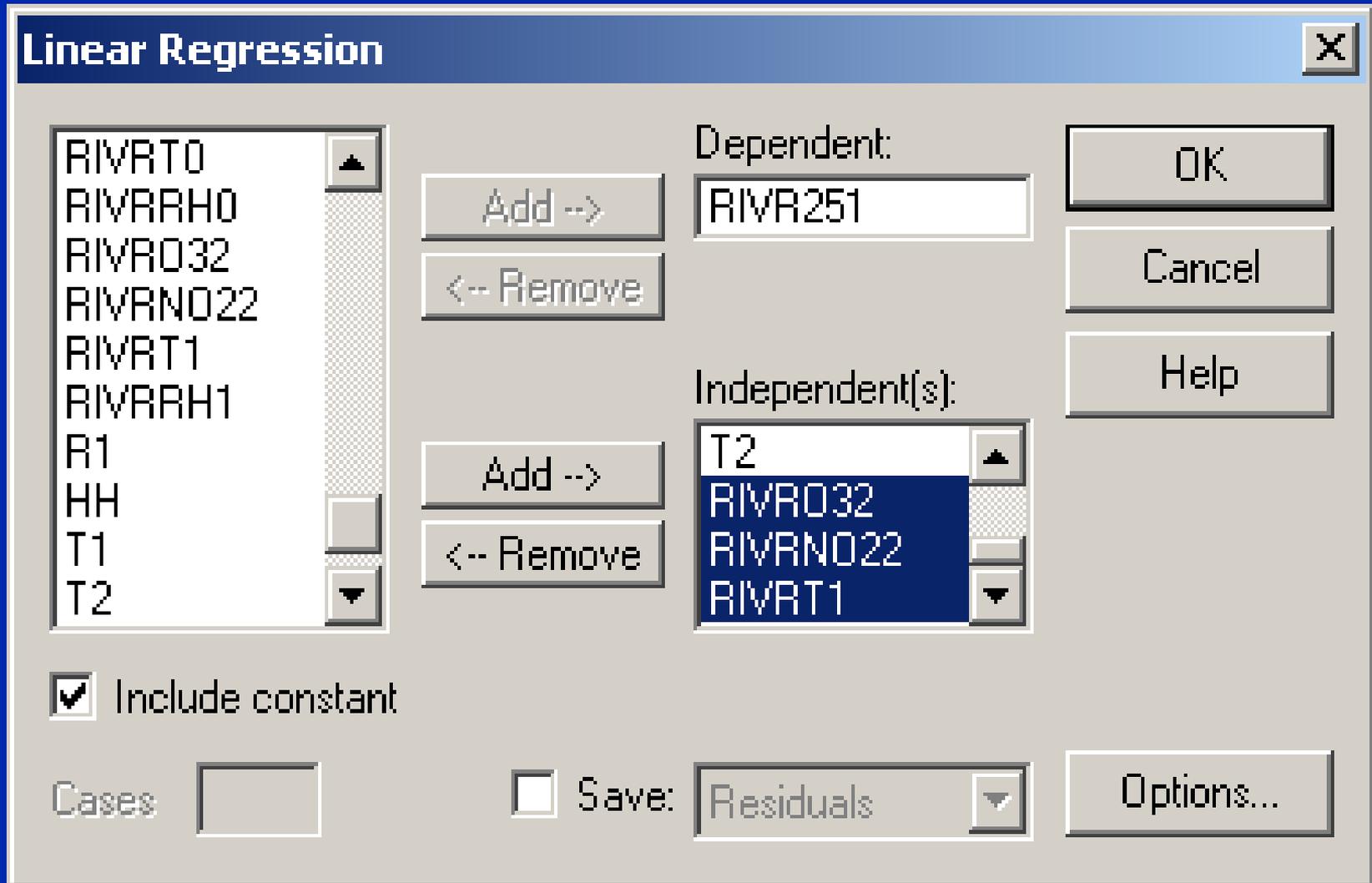
AND =

AND =

Complete Turn Select Off

OK Cancel Help

SYSTAT – Multivariate Regression



SYSTAT – Adding Stepwise Option

Regression: Options [X]

Tolerance:

Estimation

Complete

Mixture model

Stepwise

Stepwise Options

Backward Automatic

Forward Interactive

Probability

F-Statistic

Force: Max step:

SYSTAT – Regression Equation, Fit, and Statistics

Untitled - SYSTAT Output Organizer

File Edit View Data Graph Statistics BMDP Help

Courier New 9 B I U

Statistics - Statistics

Effect	Coefficient	Std Error	Std Coef	Tolerance	t
CONSTANT	-60.216	26.111	0.000	.	
HOURS	-3.111	0.461	-0.312		0.252
H8VBG	0.634	0.162	0.140		0.420
T8VBG	0.512	0.144	0.240		0.119
SUMPG	1.568	0.249	0.978		0.022
LAXSFO	-0.782	0.164	-0.137		0.654
SANLAS	-3.379	0.532	-0.925		0.025
T1000	-1.265	0.244	-0.240		0.251
T950	-0.626	0.207	-0.208		0.114
RH950	0.130	0.028	0.229		0.215
RH800	-0.150	0.019	-0.250		0.511
RIVR032	0.503	0.246	0.087		0.297
RIVR022	3.623	0.541	0.223		0.484
RIVRRH1	0.237	0.035	0.267		0.346
R1	0.105	0.034	0.139		0.264
T2	0.471	0.115	0.217		0.193
UCRIT2	0.117	0.033	0.124		0.450

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio
Regression	122309.845	16	7644.365	6
Residual	94708.825	810	116.924	

*** DELETED ***
 (YEAR= -1)
 109 case(s) deleted due to missing data.

Dep Var: RIVER251 N: 827 Multiple R: 0.751 Squared multiple R: 0.564
 Adjusted squared multiple R: 0.555 Standard error of estimate: 10.813

Ready NUM

Interpreting the Equation

Variable	Coefficient/Intercept	Contribution to PM _{2.5} Formation Process
CONSTANT	-60.216	
HOURS	-3.111	Compensates for summer insolation
H8VBG	0.634	Stagnation
T8VBG	0.512	Stagnation
SUMPG	1.568	Westerly onshore flow
LAXSFO	-0.782	Southerly onshore flow
SANLAS	-3.379	Competing with SUMPG
T1000	-1.265	Warm temps greater mixing
T950	-0.626	Competing T8VBG or RH1
RH950	0.130	Stratus or saturated layer
RH850	-0.150	Very deep mixing
RIVRO32	0.503	Summer ozone (up to 10 µg/m ³)
RIVRNO22	3.623	Fall NO ₂ contribution
RIVRRH1	0.237	Surface heterogeneous chemistry
R1	0.105	Surface heterogeneous chemistry
T2	0.471	Photochemistry indicator
UCRIT2	0.117	PM ₁₀ trend contribution

Decision Tree – Options to Consider

- Variable transformations
- Case selection
- New variables
- Missing data less an issue
- Address algorithm fitting criteria

SYSTAT – Growing the Tree

Trees [X]

YEAR
MONTH
DAY
DWK0
DWK1
P0
P1
HOURS
H5VBG
H5SAN

Add -->
<-- Remove

Add -->
<-- Remove

Dependent:
RIVR251

Independent(s):
SANLAS
T950
T900
T850

Expand model

Loss: Least squares

Display nodes as: Text

OK
Cancel
Help

Stopping...

SYSTAT – Tree Statistics, Splits, and Nodes

Untitled - SYSTAT Output Organizer

File Edit View Data Graph Statistics BMDP Help

Courier New 9 B I U

data for the following results were selected according to:
(YEAR> -1)
789 cases deleted due to missing data.

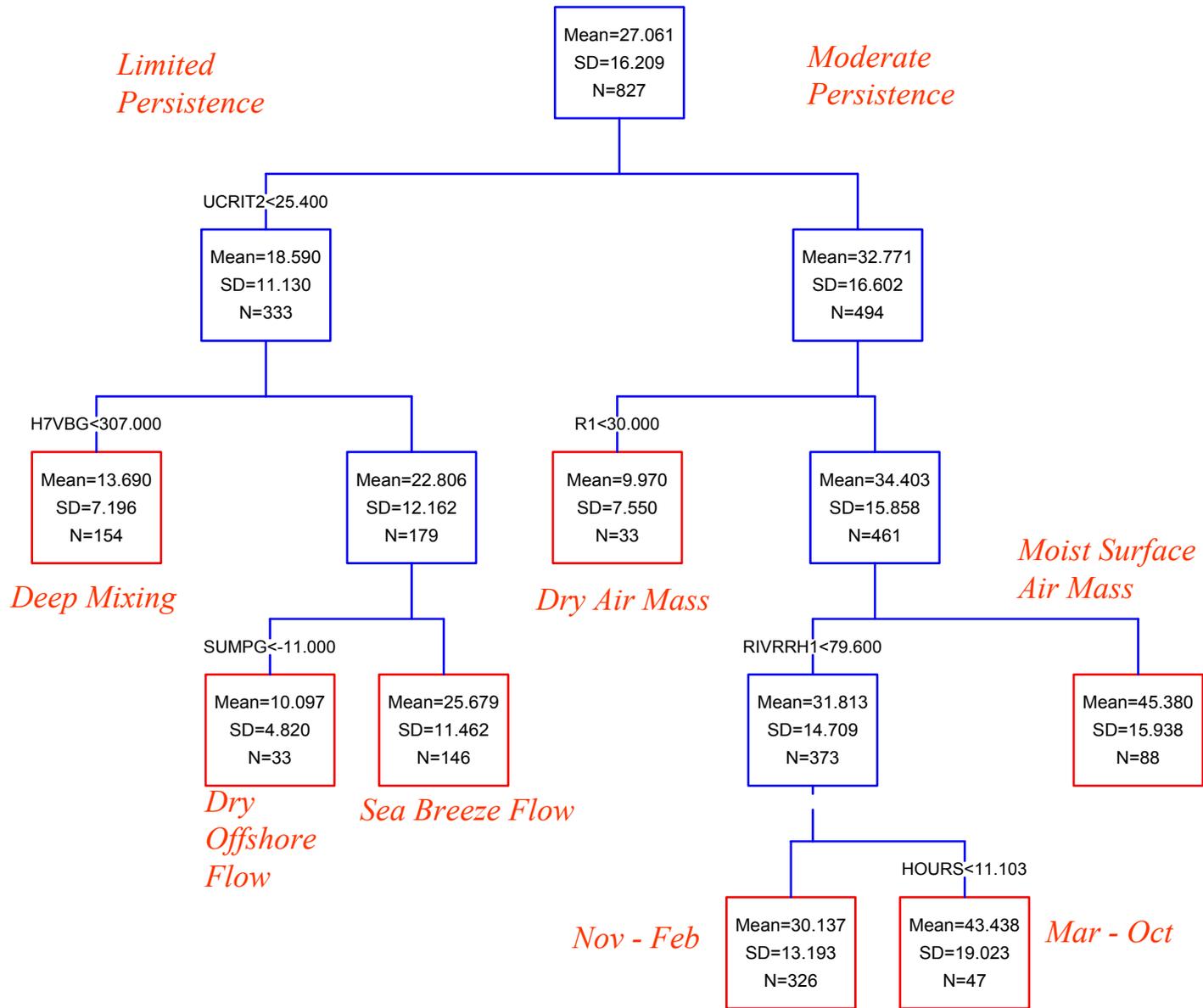
Split	Variable	PRE	Improvement
1	UCRIT2	0.184	0.184
2	H7VBG	0.216	0.032
3	SUMPG	0.246	0.030
4	R1	0.331	0.085
5	RIVRRH1	0.391	0.060
6	HOURS	0.425	0.033

Fitting Method: Least Squares
Predicted variable: RIVR251
Minimum split index value: 0.030
Minimum improvement in PRE: 0.030
Maximum number of nodes allowed: 22
Minimum count allowed in each node: 5
The final tree contains 7 terminal nodes
Proportional reduction in error: 0.425

Node	from	Count	Mean	SD	Split Var	Cut Value	Fit
1	0	827	27.061	16.209	UCRIT2	25.400	0.184
2	1	333	18.590	11.130	H7VBG	307.000	0.167
3	1	494	32.771	16.602	R1	30.000	0.135
4	2	154	13.690	7.196			
5	2	179	22.806	12.162	SUMPG	-11.000	0.248
6	5	33	10.097	4.820			
7	5	146	25.679	11.462			
8	3	33	9.970	7.550			
9	3	461	34.403	15.858	RIVRRH1	79.600	0.113
10	9	373	31.813	14.709	HOURS	11.103	0.090
11	9	88	45.380	15.938			
12	10	47	43.438	19.023			
13	10	326	30.137	13.193			

Ready NUM

RIVR251



Understanding the Tree

- Do the splits make physical sense?
- Does the structure fit the conceptual model?
- Can the tree be further pruned?
- Does the tree need to be grown?
- Can the tree be used directly as a forecast tool?

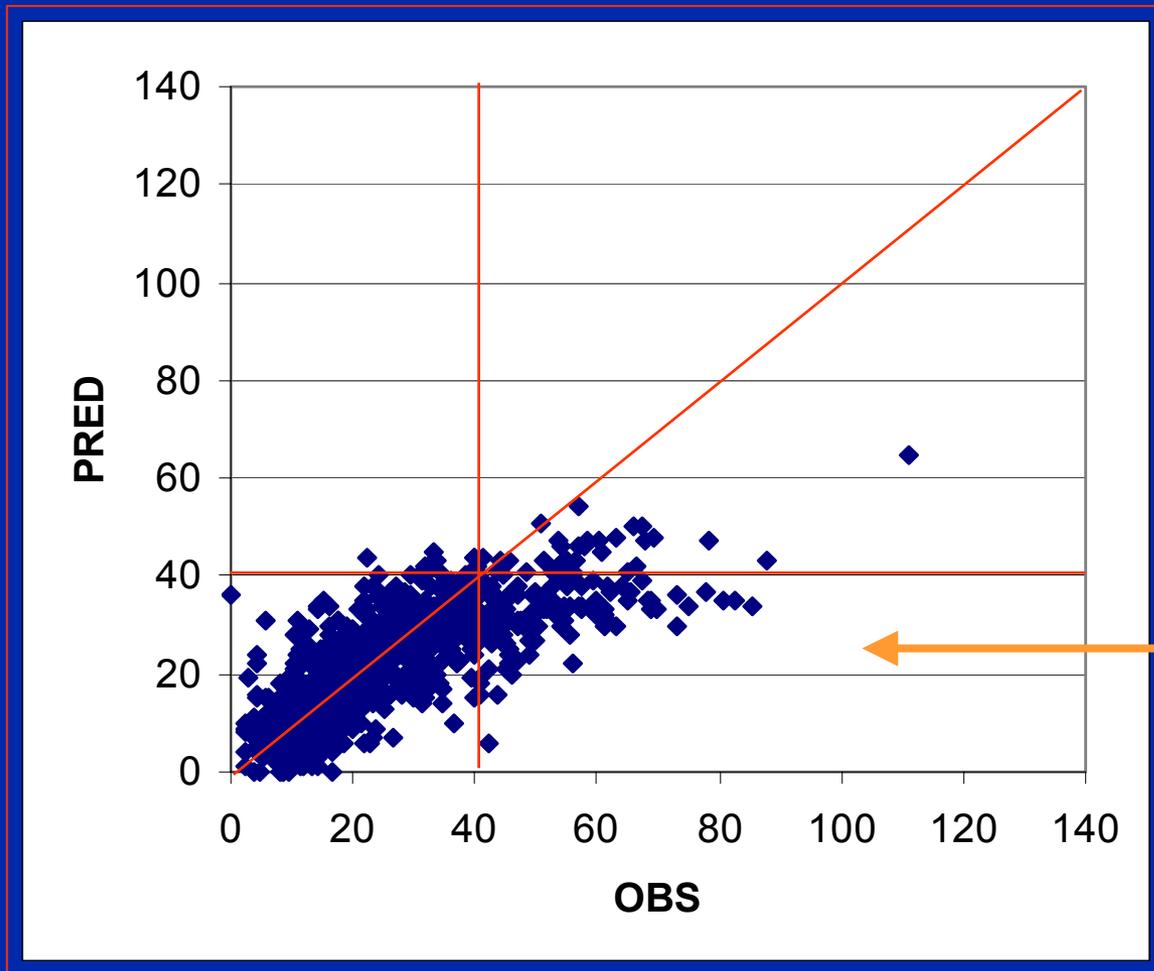
Model Evaluation – Evaluating the Branching and Terminal Nodes

Node	PM2.5 µg/m3	AQI	Percent Distribution	Description
1	13.7	G	18.6	Deep mixing with limited persistence
2	10.1	G	4.0	Dry offshore flow with limited persistence (Santa Ana)
3	25.7	M	1.7	Onshore flow with moderate mixing and limited persistence
4	10.0	G	4.0	Dry offshore low with moderate persistence
5	31.1	M	39.4	Nov-Feb heterogeneous nitrate chemistry
6	43.4	US	5.7	Mar-Oct combined photochemical and heterogeneous chemistry
7	45.3	US	10.6	Stagnant air mass with fog

Model Evaluation – General

- Assess performance using development and independent data sets
- Set a target for model performance
- Contingence matrix for threshold analyses (>50%)
- Error (10% of maximum observed concentration)
- Bias as near to zero as desired

Model Evaluation – PM_{2.5} Regression Model Performance



Model tendency:
underpredict

Alternative Models

- Combinations of CART and regression
- Discriminate or cluster analyses to group similar air quality events
- Factor analysis
- Nearest neighbor
- Neural networks

Daily Operation

- The model/algorithm is a tool!
- Bad or missing input data will be reflected in the prediction
- Adjustments are required for special circumstances and timing
- Forecasters blend experience with model output to generate the final product
- Have fun!