



UERRA General Assembly 3

WP1 preliminary quality control results

Linden Ashcroft, Manola Brunet, Mercè Castella and Joan Ramon Coll

Toulouse 1 February 2016

1. Digitisation

- Template preparation
- Self-checking by digitisers



2. Visual Cross Checking

- Systematic visual checking
- Unit identification



3. Database ingestion

- Date and non-numeric errors removed
- Data standardised



4. Automatic quality control

- 14 tests applied
- Results manually examined

1. Digitisation

- Template preparation
- Self-checking by digitisers

2. Visual Cross Checking

- Systematic visual checking
- Unit identification

3. Database ingestion

- Date and non-numeric errors removed
- Data standardised

4. Automatic quality control

- 14 tests applied
- Results manually examined

INSTITUTO NAL. DE METEOROLOGIA
BANCO DE DATOS

TEMPERATURAS

PAG. 373
AÑO 1977 MES 1

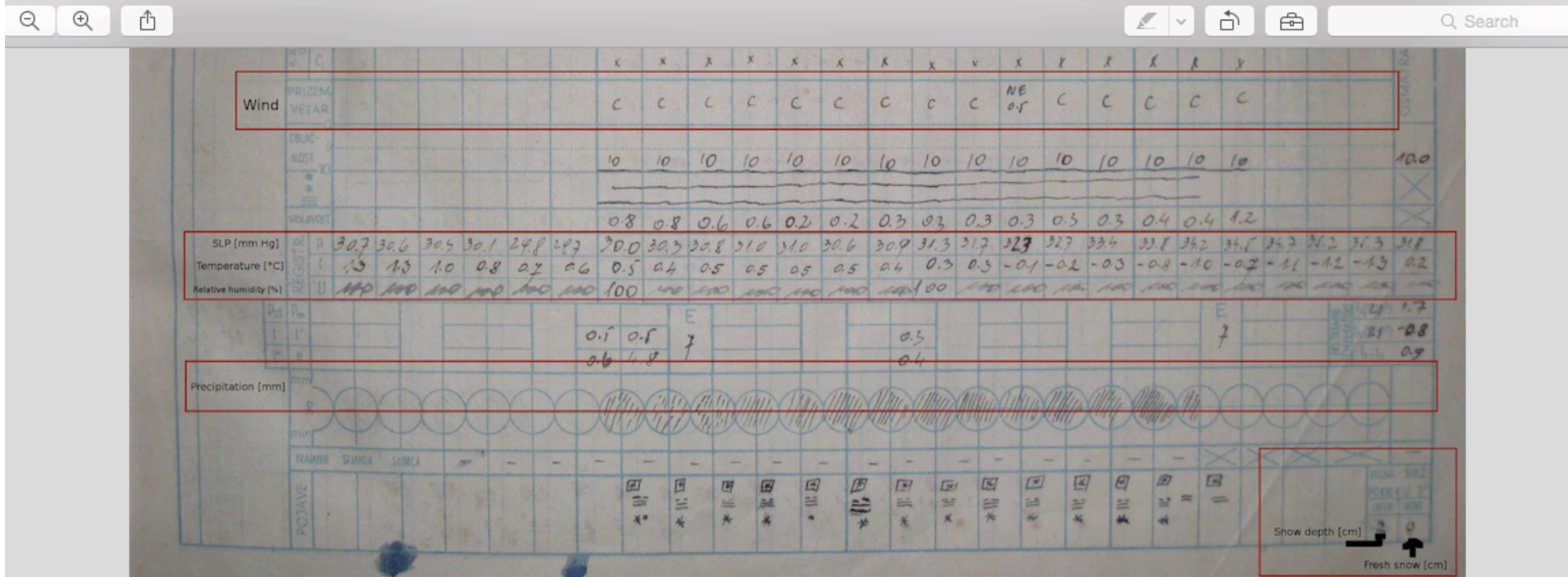
ESTACION 0042 TARRAGONA

TARRAGONA LONG. 0456E; LATI. 4106; ALTIT. 58; ALTIBA. ; (EDICION 14-07-89)

00 HORAS						07 HORAS					13 HORAS					18 HORAS					EXTREMAS					
DIA	TS	TH	HU	TV	PR	TS	TH	HU	TV	PR	TS	TH	HU	TV	PR	TS	TH	HU	TV	PR	MAX.	HORA	MIN.	HORA	MEDIA	EVP.
1						12.0	12.0	00	14.0	12.0	15.0	10.0	49	8.3	4.3	13.2	9.0	54	8.1	4.1	15.6	14.00	5.6	06.00	10.6	3.6
2						10.0	7.6	70	8.5	4.7	12.0	10.0	76	10.7	8.0	11.0	9.0	76	9.9	6.9	13.0	14.00	9.0	06.00	11.0	2.8
3						7.8	7.2	92	9.7	6.6	8.0	7.6	95	10.1	7.2	9.0	7.0	74	8.4	4.6	11.6	15.00	7.6	06.00	9.6	0.8
4						3.0	2.0	83	6.3	0.4	8.0	5.0	59	6.3	0.6	7.6	4.0	51	5.2	-1.8	9.0	14.00	2.6	06.00	5.8	1.5
5						3.4	1.2	64	4.9	-2.8	4.6	3.0	75	6.3	0.5	5.0	3.0	69	6.0	-0.1	7.0	15.00	0.2	00.30	3.6	1.3
6						5.6	4.0	76	6.8	1.7	11.4	9.8	81	10.8	8.2	10.0	9.0	87	10.7	8.0	11.6	16.30	3.4	06.00	7.5	4.5
7						11.8	11.0	91	12.5	10.3	12.4	10.6	79	11.3	8.9	13.0	11.0	77	11.5	9.1	15.0	14.00	5.2	06.00	10.1	3.3
8						9.2	8.0	84	9.8	6.7	13.0	10.0	66	9.9	6.9	12.0	9.6	72	10.0	7.1	14.0	15.00	8.0	06.00	11.0	2.6
9						4.0	2.6	77	6.2	0.4	8.8	5.2	53	6.0	-0.1	9.0	5.0	49	5.5	-1.2	11.6	14.00	3.4	06.00	7.5	4.0
10						3.6	2.0	74	5.8	-0.6	9.4	6.8	67	7.8	3.5	9.0	7.0	74	8.4	4.6	11.0	14.00	2.2	03.00	6.6	2.8
10.						7.0	5.8	81	8.4	3.9	10.3	7.8	70	8.7	4.8	9.9	7.4	68	8.4	4.1	11.9		4.7		8.3	2.7
11						7.0	5.4	77	7.7	3.3	7.4	6.0	80	8.2	4.2	6.6	4.0	63	6.0	0.0	8.0	12.00	2.4	06.00	5.2	4.0
12						5.4	2.8	61	5.4	-1.5	9.0	5.0	49	5.5	-1.2	8.2	4.6	52	5.6	-1.0	10.0	15.00	3.8	01.00	6.9	14.0
13						6.4	4.2	68	6.5	0.9	11.0	7.0	52	6.8	1.6	11.6	9.0	69	9.4	6.1	13.4	13.30	4.4	06.30	8.9	6.6

Tarr_Temp_template.xls

Search in Sheet																				
Home Layout Tables Charts SmartArt Formulas Data Review Developer																				
Q1																				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Year	Month	Day	0700 TD		0700 HU		0700 PR	1300 TD		1300 HU		1300 PR	1800 TD		1800 HU		1800 PR		
2	1977	1	1																	
3	1977	1	2																	
4	1977	1	3																	
5	1977	1	4																	
6	1977	1	5																	
7	1977	1	6																	
8	1977	1	7																	
9	1977	1	8																	
10	1977	1	9																	
11	1977	1	10																	
12																				
13	1977	1	11																	
14	1977	1	12																	
15	1977	1	13																	
16	1977	1	14																	
17	1977	1	15																	
18	1977	1	16																	
19																				
Sheet1																				
Normal View Ready Sum=0																				



HomeLayoutTablesChartsSmartArtFormulasDataReviewDeveloper

A5842

<

Initial quality control at digitisation stage

- Students should check each month of data
- Templates used for 30% of sources
- Mixed responses to template use from digitizers

1. Digitisation

- Template preparation
- Self-checking by digitisers

2. Visual Cross Checking

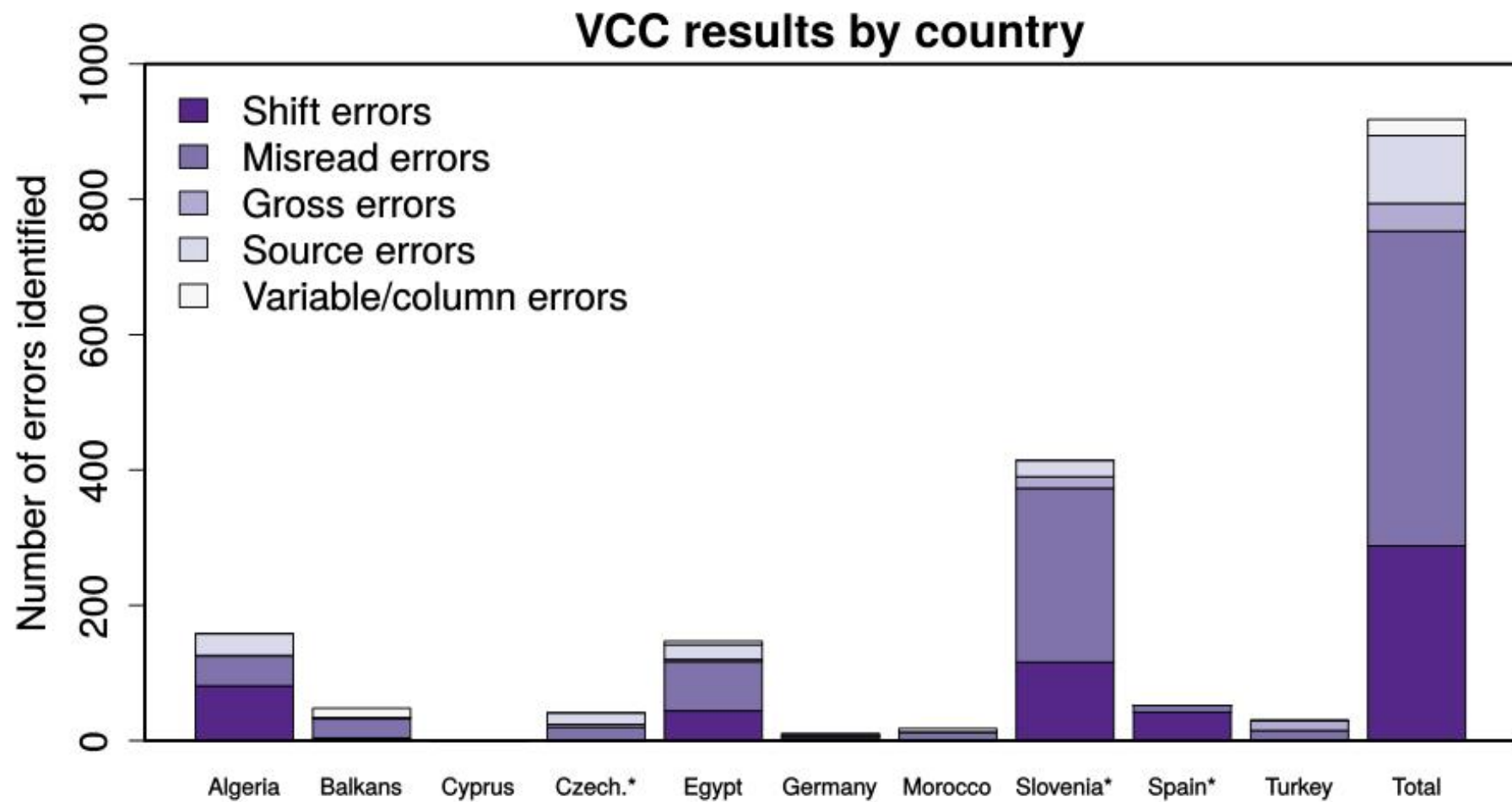
- Systematic visual checking
- Unit identification

3. Database ingestion

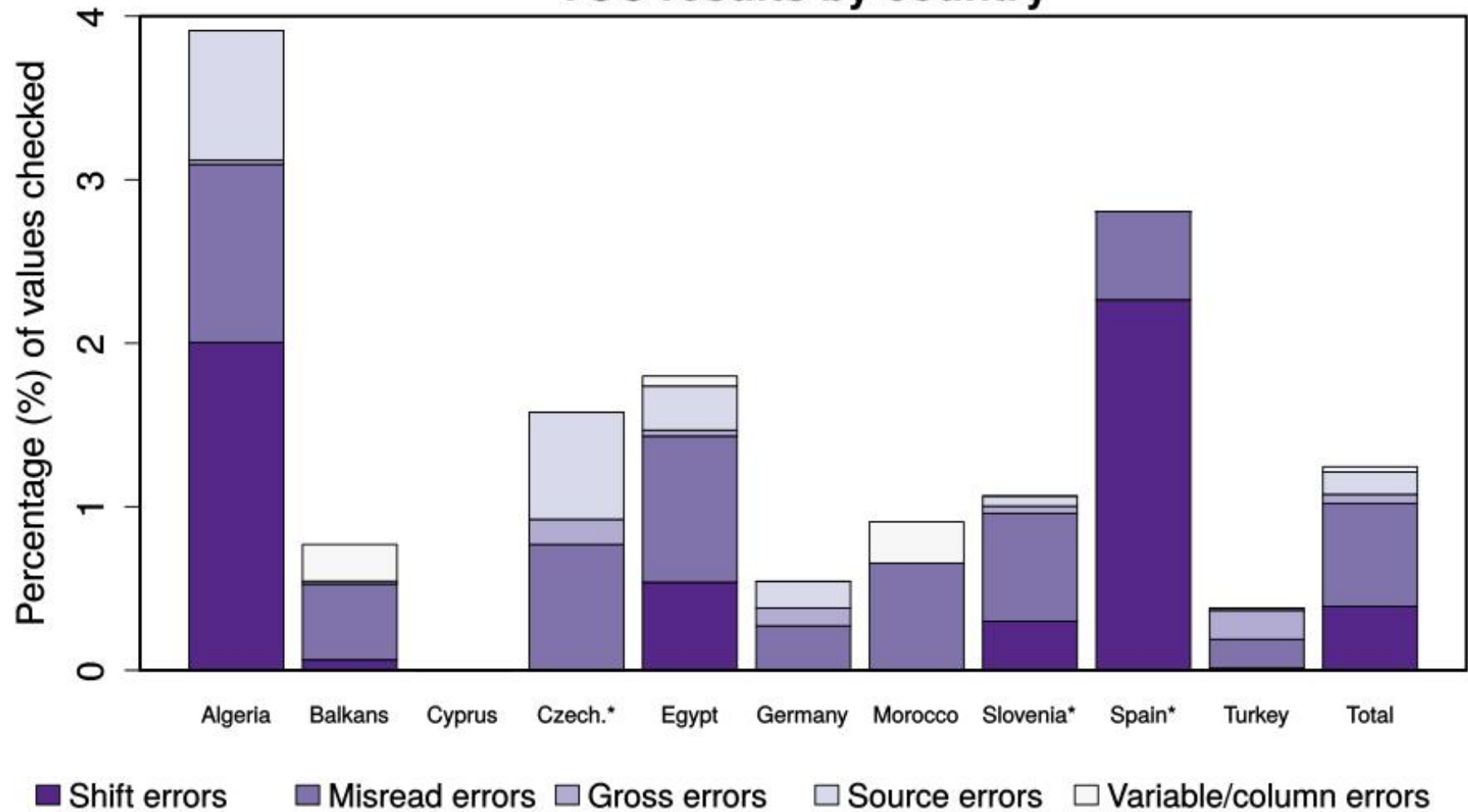
- Date and non-numeric errors removed
- Data standardised

4. Automatic quality control

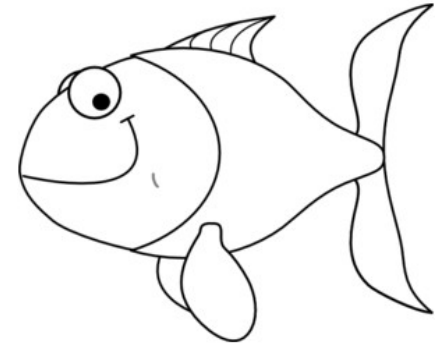
- 14 tests applied
- Results manually examined



VCC results by country



Other big fish found



- Instructions misunderstood
- Incorrect variables digitised
- Zero rainfall and snowfall recorded as missing
- But overall, not bad...

1. Digitisation

- Template preparation
- Self-checking by digitisers

2. Visual Cross Checking

- Systematic visual checking
- Unit identification

3. Database ingestion

- Date and non-numeric errors removed
- Data standardised

4. Automatic quality control

- 14 tests applied
- Results manually examined

Database ingestion and automatic quality control preparation

- Invalid dates, non-numeric values identified
- Data now accessible in standardised format



Unit standardisation, aka, the joys of data

Station level pressure and sea level pressure in the same source

STATIONS	ALTITUDE (m) (de la cuvette du baromètre)	OBSERVATIONS	
		Pression barométrique à 0 degré C et au niveau de la mer en millibars et dixièmes de millibars	Pression au niveau de la mer
Tanger.	86	1012.8	
Oujda	460		
Port-Lyautey	6	1011.1	
Touahar	569	1011.6	
Rabat	68	1012.1	
Fès	414	1011.5	
Meknès	549	1010.6	
Nouasseur	200	1013.2	
Casablanca	85	1012.4	
Ifrane (M)	1636	839.1	
Khouribga (M)	776	924.7	
Safi	25	1010.3	
Kasba-Tadla	501		
Midelt (M)	1520	847.1	
Ksar es souk (M)	1060		
Mogador	8	1010.1	
Marrakech	466	1010.3	
Oukaimeden (M)	2639		
Agadir	19	1009.9	
Ouarzazate (M)	1133	887.7	
Tindouf	451		
Fort-Trinquet.	360	1009.4	

Wind direction and strength changing over time

Bofor No.	İsimler	H I Z	
		10 m. yükseklikte (Saniyede metre)	
0	Sakin	0.0	0.2
1	Hafif esinti		
2	Hafif briz	0.3	1.5
3	Zayıf briz	1.6	3.3
4	Mutedil briz	3.4	5.4
5	Sert briz	5.5	7.9
6	Kuvvetli Rüzgâr	8.0	10.7
7	Şiddetli rüzgâr	10.8	13.8
8	Fırtınamsı rüzgâr	13.9	17.1
9	Fırtına	17.2	20.7
10	Şiddetli fırtına	20.8	24.4
11	Orkanımsı fırtına	24.5	28.4
12	Orkan - Kasırga	28.5	32.6

1. Digitisation

- Template preparation
- Self-checking by digitisers



2. Visual Cross Checking

- Systematic visual checking
- Unit identification



3. Database ingestion

- Date and non-numeric errors removed
- Data standardised



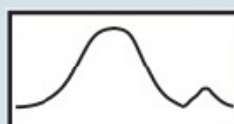
4. Automatic quality control

- 14 tests applied
- Results manually examined

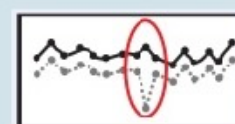
Automatic quality control (AQC) tests



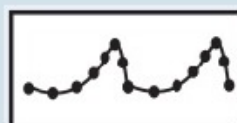
Date order



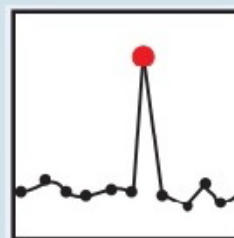
Strange distributions



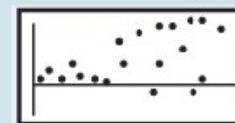
Calculated vs observed values



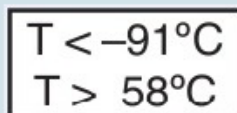
Pattern repetition



Climatic outliers



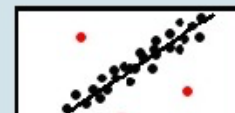
Strange scattering



Record breakers



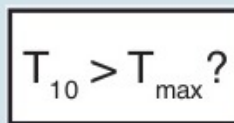
Jumps and spikes



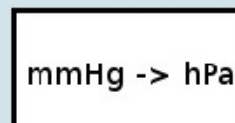
Bivariate distribution outliers



Repeated values



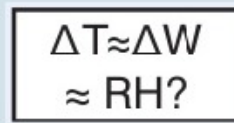
Logical failures



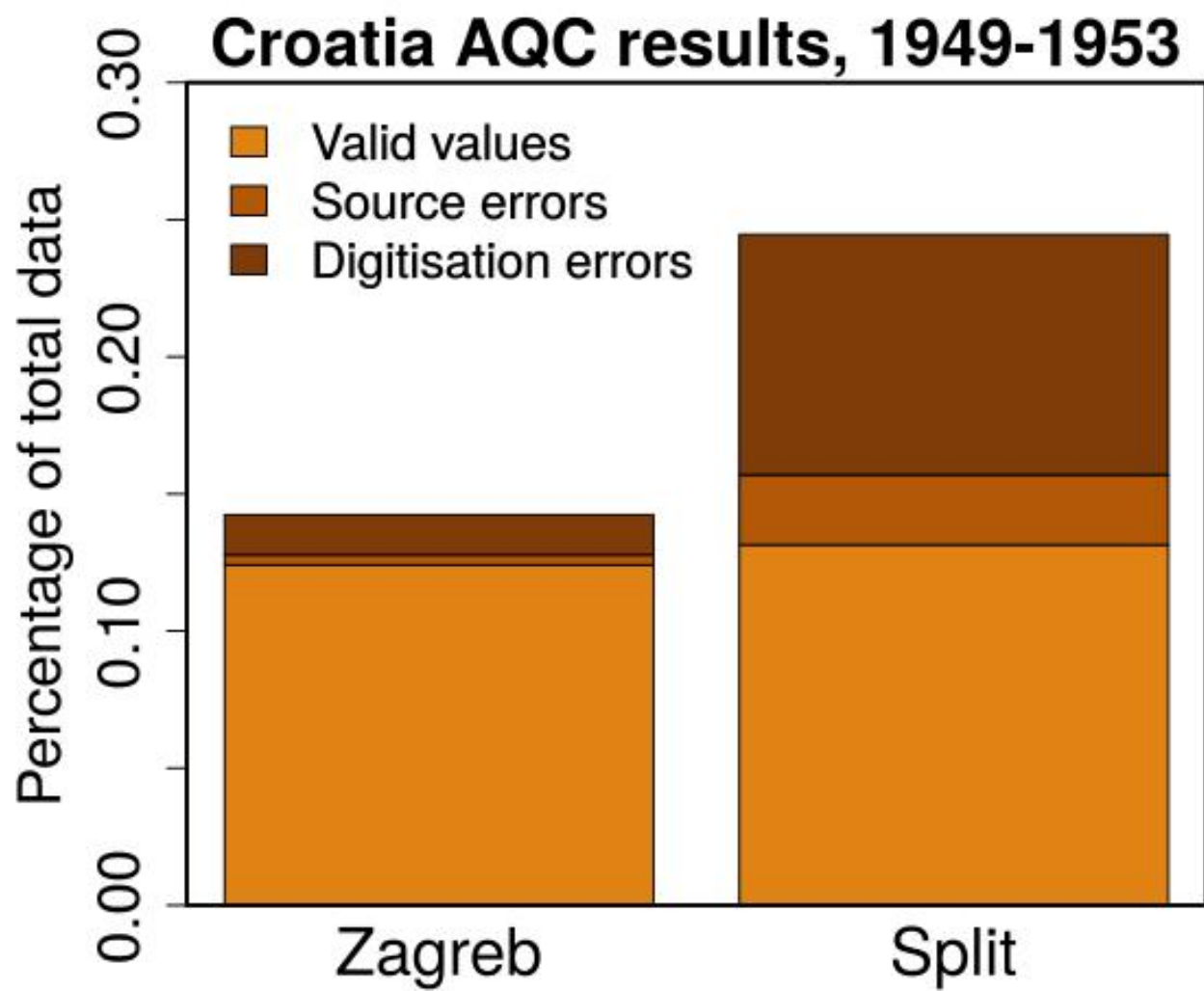
Unit changes

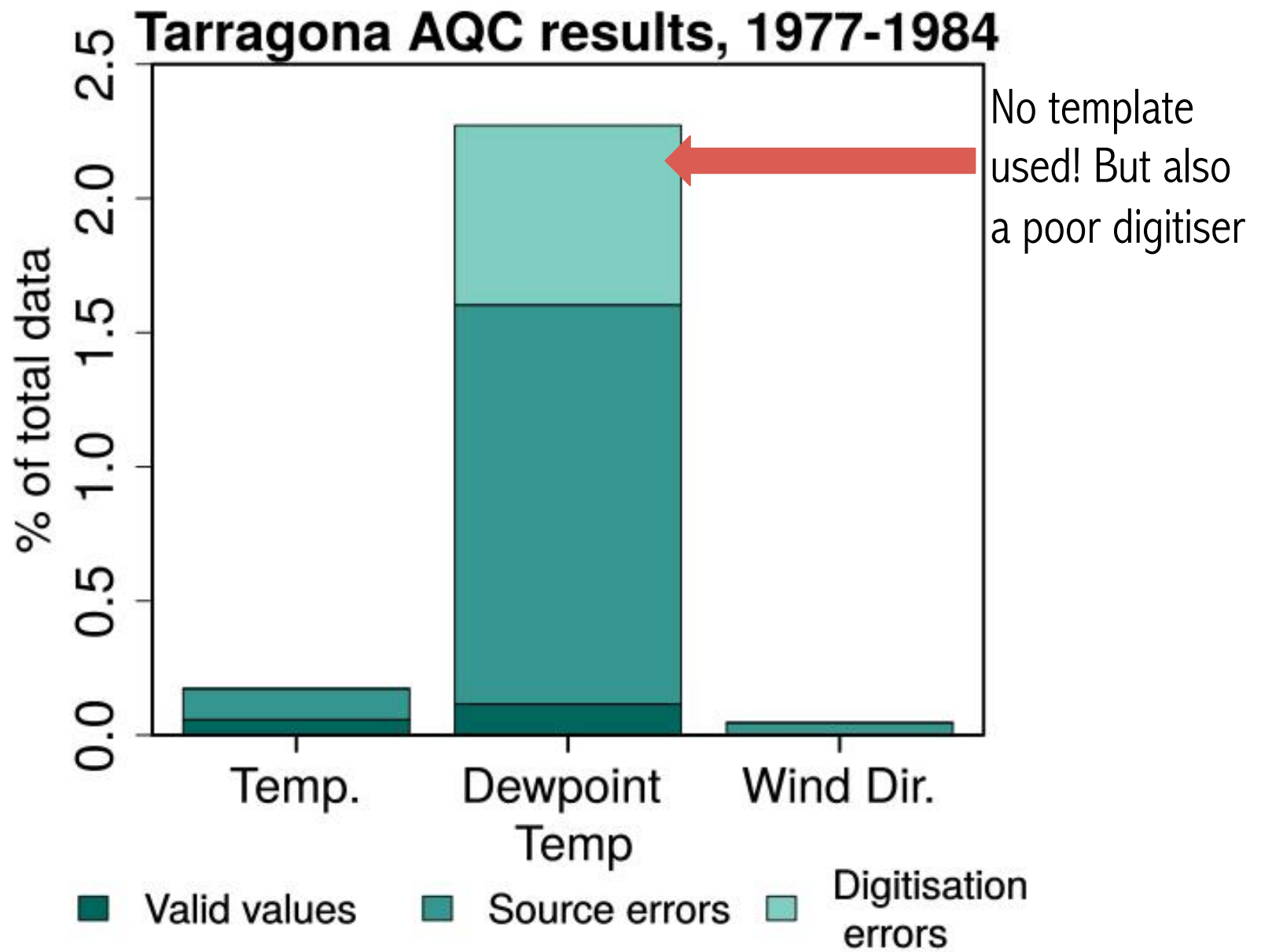


Frequency biases



Intervariable comparison





AQC improvements to be made

- Reduce sensitivity of bivariate checks for pressure
- Reassess quality checking of categorical wind observations

Summary

- Visual cross checking gives a good indication of data quality
- Digitised data reliability depends on digitiser, data source complexity and use of templates
- Automatic quality control can identify digitisation errors but some tests are too sensitive and wind values are difficult

lindenclaire.ashcroft@urv.cat | [@lindenashcroft](https://twitter.com/lindenashcroft)