Information Extraction from Handwritten Tables in Historical Documents

José Andrés[1], Jose Ramón Prieto[1], Emilio Granell[1], Verónica Romero[2], Joan Andreu Sánchez[1], Enrique Vidal[1]

[1]Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain

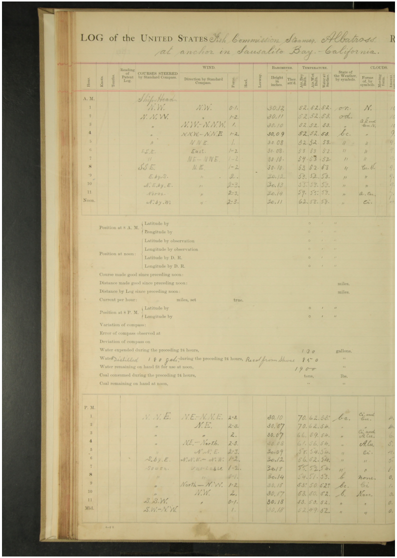
[2]Departament d'Informàtica, Universitat de València, Spain.

May 23rd, 2022

PRHLT

Introduction

Introduction



LOG of the UNITED STATES Fish Commission Steamer. Albatross. R
at anchor in Sausalito Bay.-California.

Hour.	Knots.	Tenths.	Reading of Patent Log.	COURSES STEERED by Standard Compass.	WIND.		Heel.	Leeway.	BAROMETER.		TEMPERATURE.			State of the Weather, by symbols.	CLOUDS.		
					Direction by Standard Compass.	Force.			Height in inches.	Ther. att'd.	Air, Dry Bulb.	Air, Wet Bulb.	Water at Surface.		Forms of, by symbols.	Moving form.	Amount
A. M.				Ship Head													
1				N.W.	N.W.	0-1.			30.12		52.	52.	52.	or.	N.		10
2				N.N.W.	"	1-2.			30.11		52.	52.	53.	od.	"		10
3				"	N.W.-N.N.W.	1.			30.10		52.	52.	53.	"	A.S. and Cu.N.		10
4				"	N.N.W.-N.N.E	1-2.			30.09		52.	52.	53.	bc.	"		9.
5				"	N N E.	1.			30.08		52	52	53.	"	"		9.
6				E.S.E.	East.	1-2			30.08.		53	53	52.	"	"		4.
7				"	NE-NNE.	1-2			30.10.		54.	53	52	"	"		9.
8				SSE.	N.E.	1-2			30.10.		53	52.	53	"	Cu.N.		9.
9				E.by.S.	"	2.			30.12		53.	52.	53.	"	"		9.
10				N.E.by.E.	"	2-3.			30.13		53.	53.	53.	"	"		9.
11				North.	"	2-3.			30.14		59.	55.	53.	"	a. Cu.		7.
Noon.				N.by.W.	"	2-3.			30.11		62.	58.	53.	"	Ci.		4

Position at 8 A. M. { Latitude by ° ′ ″
{ Longitude by ° ′ ″

Position at noon: { Latitude by observation ° ′ ″
Longitude by observation ° ′ ″
Latitude by D. R. ° ′ ″
Longitude by D. R. ° ′ ″

Course made good since preceding noon:

Distance made good since preceding noon: miles.

Distance by Log since preceding noon: miles.

Current per hour: miles, set true.

Position at 8 P. M. { Latitude by ° ′ ″
{ Longitude by ° ′ ″

Variation of compass:

Error of compass observed at

Deviation of compass on

Water expended during the preceding 24 hours, 130 gallons.

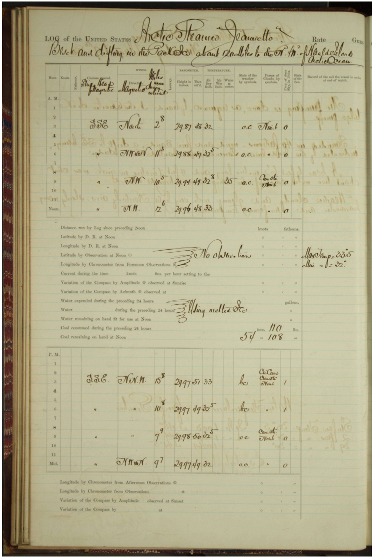
Water Distilled 180 gals; during the preceding 24 hours, Recd from Shore 850 "

Water remaining on hand fit for use at noon, 1900 "

Coal consumed during the preceding 24 hours, tons, lbs.

Coal remaining on hand at noon, " "

Hour.	Knots.	Tenths.	Reading of Patent Log.	COURSES STEERED by Standard Compass.	Direction by Standard Compass.	Force.	Heel.	Leeway.	Height in inches.	Ther. att'd.	Air, Dry Bulb.	Air, Wet Bulb.	Water at Surface.	State of the Weather, by symbols.	Forms of, by symbols.	Moving form.	Amount
P. M.																	
1				N. N. E.	N.E.-N.N.E.	2-3.			30.10		70.	62.	53.	bc.	Ci. and Cu.		4.
2				"	N.E.	2-3.			30.07		70.	62.	54.	"	"		4.
3				"	"	2.			30.07		66.	59.	54.	"	Ci. and A.Cu.		6.
4				"	N.E.-North.	2-3.			30.08		61.	56.	54.	"	A.Cu.		5.
5				"	N.N.E.	2-3.			30.09		58.	54.	54.	"	Ci.		4.
6				S.by.E.	N.N.E.-N.W.	1-2.			30.12		56.	52.	54.	"	"		5.
7				South.	Variable	1-2.			30.13		55.	52.	54.	"	"		1.
8				"	"	0-1.			30.14		54.	51.	53.	b	none.		0.
9				"	North-N.W.	1-2.			30.15		53.5	50.	52.5	bc.	Ci		1.
10				"	N.W.	2.			30.17		53.	50.	52.	b.	None.		0.
11				S.S.W.	"	0-1.			30.18		53.	50.	52.	"	"		0.
Mid.				S.W.-N.W.	"	1.			30.18		52.	49.	52.	"	"		0.



LOG of the UNITED STATES Arctic Steamer Jeannette, Rate Guns
Beset and drifting in the Pack Ice about 120 Miles to the N.W. of Herald Island Arctic Ocean.

Hour.	Knots.	Fathoms.	Courses steered. Ship Head Magnetic	WINDS. Direction. Magnetic	Miles during interval	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1																
2																
3			SSE	North	2^{8}		29.87	48	32			o.c.	Nimb	0		
4																
5																
6			"	NWbN	11^{3}		29.88	47	32^{5}			o.c.	"	0		
7																
8																
9			"	NW	10^{5}		29.94	49	32^{8}		35	o.c.	Cum.str. Nimb.	0		
10																
11																
Noon.			"	N.W.	12^{6}		29.96	48	33			o.c.	"	0		

Distance run by Log since preceding Noon knots fathoms.

Latitude by D. R. at Noon ° ′ ″

Longitude by D. R. at Noon ° ′ ″

Latitude by Observation at Noon ⊙ No observations ° ′ ″ Max Temp. = 33°.5

Longitude by Chronometer from Forenoon Observations ⊙ ° ′ ″ Min " " = 32°

Current during the time knots fms. per hour setting to the

Variation of the Compass by Amplitude ⊙ observed at Sunrise ° ′ ″

Variation of the Compass by Azimuth ⊙ observed at ° ′ ″

Water expended during the preceding 24 hours gallons.

Water during the preceding 24 hours Using melted Ice "

Water remaining on hand fit for use at Noon "

Coal consumed during the preceding 24 hours tons. 110 lbs.

Coal remaining on hand at Noon 54 " 108 "

Hour.	Knots.	Fathoms.	Courses steered.	Direction.	Miles during interval	Leeway.	Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
P. M.															
1															
2															
3			S.S.E.	N.N.W.	15^{8}		29.97	51	33			bc	Ci.Cum Cum.str Nimb.	1	
4															
5															
6			"	"	10^{8}		29.97	49	32^{5}			bc	"	1	
7															
8															
9			"	"	7^{9}		29.98	50	32^{5}			oc	Cum.str. Nimb	0	
10															
11															
Mid.			"	NWN.	9^{7}		29.97	49	32			o.c.	"	0	

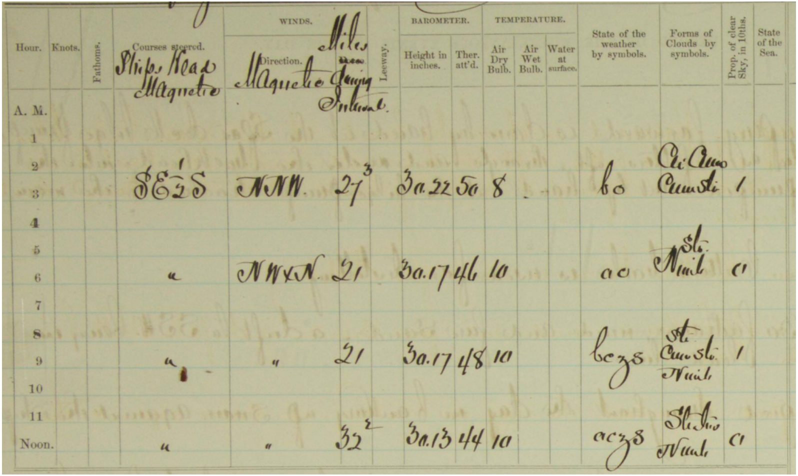
Longitude by Chronometer from Afternoon Observations ⊙ ° ″

Longitude by Chronometer from Observations * ° ′ ″

Variation of the Compass by Amplitude observed at Sunset ° ′ ″

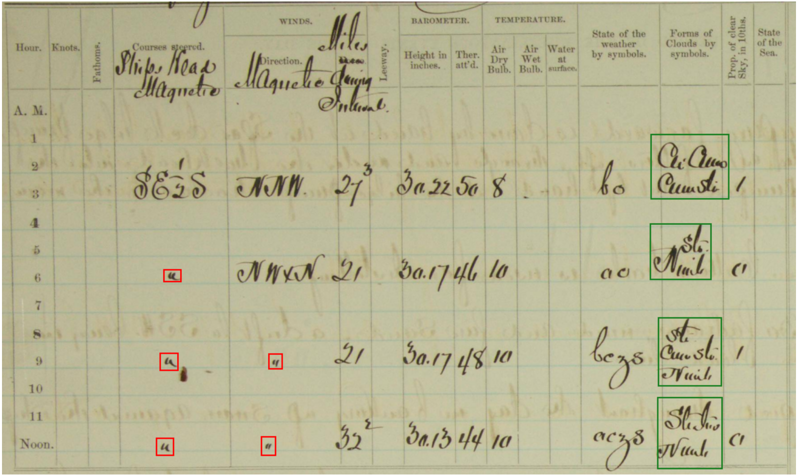
Variation of the Compass by at ° ′ ″

HisClima table example



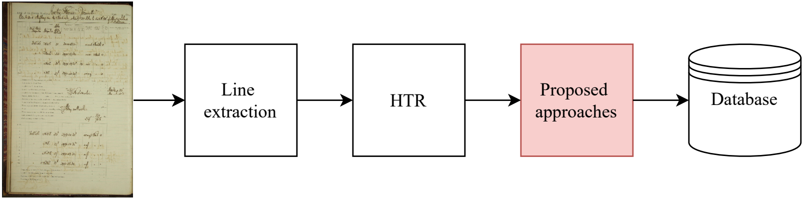
Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction. Magnetic	 Miles ~~Force~~ Gaining Interval.	Leeway.	BAROMETER. Height in inches.	 Ther. att'd.	TEMPERATURE. Air Dry Bulb.	 Air Wet Bulb.	 Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27^{3}		30.22	50	8			bo	Ci: Cum Cum Sti	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			ac	Sti. Nimb	0	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum Sti. Nimb	1	
10															
11															
Noon.			"	"	32^{2}		30.13	44	10			aczs	Sti Cu Nimb	0	

HisClima table: challenges



Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction. Magnetic	WINDS. Miles ~~Force.~~ during Interval.	Leeway.	BAROMETER. Height in inches.	BAROMETER. Ther. att'd.	TEMPERATURE. Air Dry Bulb.	TEMPERATURE. Air Wet Bulb.	TEMPERATURE. Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27³		30.22	50	8			bo	Ci: Cum Cum Str.	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			ao	St. Nimb.	0	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum St. Nimb.	1	
10															
11															
Noon.			"	"	32²		30.13	44	10			aczs	Str. Cum Nimb.	0	

General pipeline



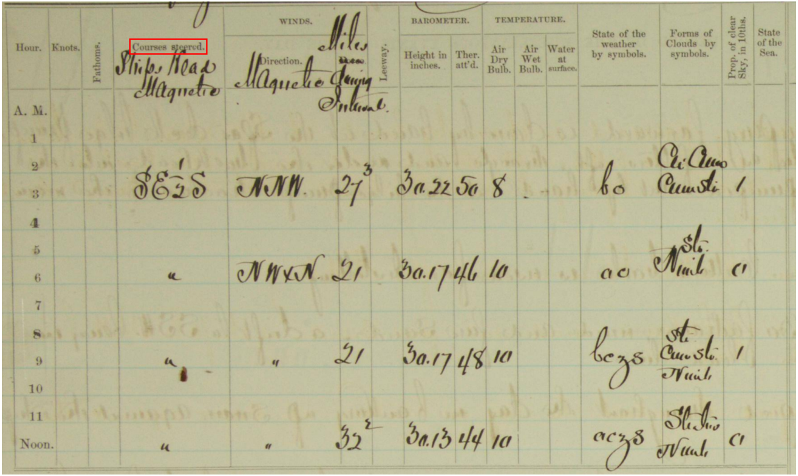
Line extraction

HTR

Proposed approaches

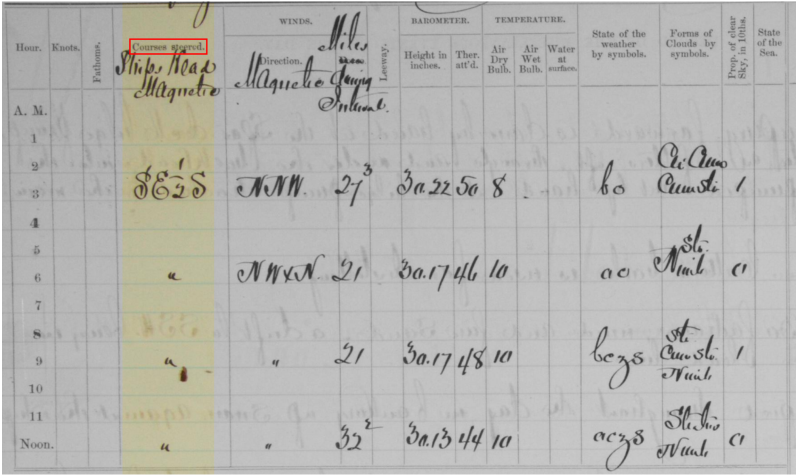
Database

Heuristic Geometric Information



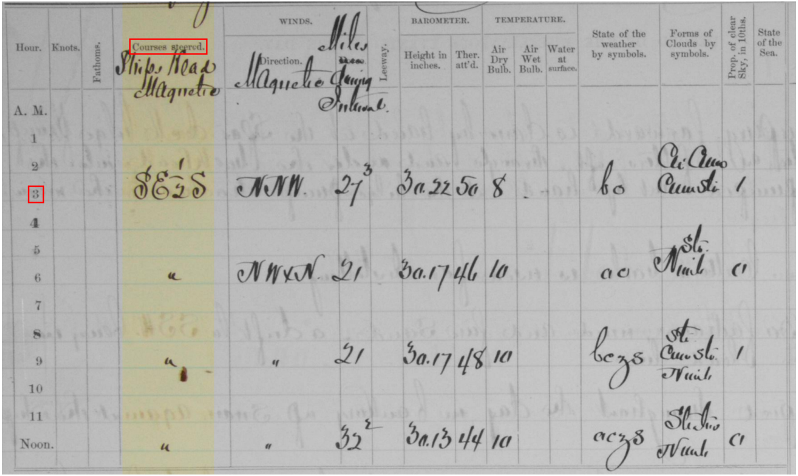
Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	WINDS. Direction. Magnetic	Miles Gauging Interval	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27^{3}		30.22	56	8			bo	Ci: Cum Cum: St.	1	
4															
5															
6			"	NWxN.	21		30.17	46	10	oc	St. Nimb	01	
7															
8															
9			"	"	21		30.17	48	10			bcqs	St. Cum: St. Nimb	1	
10															
11															
Noon.			"	"	32^{2}		30.13	44	10			ocqs	Stratus Nimb	01	

Heuristic Geometric Information



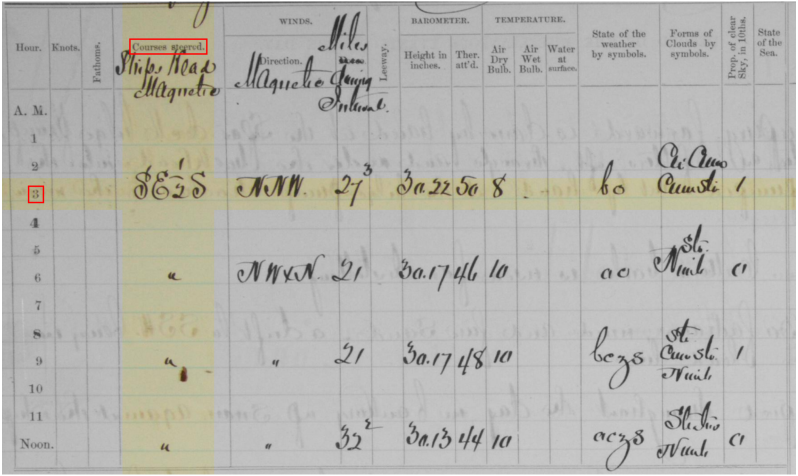
Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS.		Leeway.	BAROMETER.		TEMPERATURE.			State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
				Direction. Magnetic	~~Force.~~ Miles Gusting Interval.		Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.				
A. M.															
1															
2															
3			SE½S	NNW.	27^3		30.22	50	8	.		bo	Ci.Cum Cum.St.	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			ao	St. Nimb.	0	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum.St. Nimb.	1	
10															
11															
Noon.			"	"	32^2		30.13	44	10			aczs	St.Cu Nimb.	0	

Heuristic Geometric Information



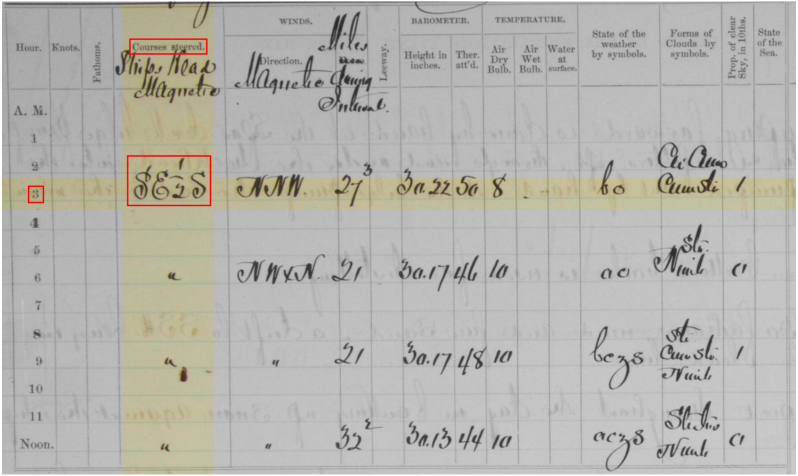
				WINDS.			BAROMETER.		TEMPERATURE.						
Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	Direction. Magnetic	Miles ~~Force.~~ during Interval.	Leeway.	Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27³		30.22	50	8	.		bo	Ci.Cum Cum.St.	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			ac	St. Nimb	c1	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum.St. Nimb.	1	
10															
11															
Noon.			"	"	32²		30.13	44	10			aczs	St.Cum. Nimb.	c1	

Heuristic Geometric Information



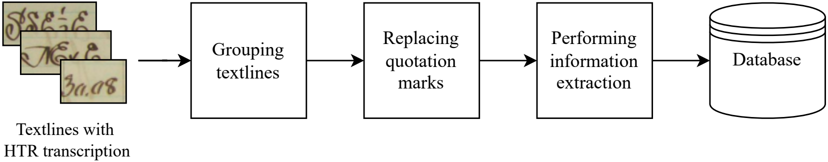
Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction. Magnetic	Miles Gauging Interval.	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2													Ci.Cum		
3			SE½S	NNW.	27^{3}		30.22	50	8	.		bo	Cum.Sti.	1	
4															
5													Sti.		
6			"	NWxN.	21		30.17	46	10			ao	Nimb.	0	
7															
8													Sti.		
9			"	"	21		30.17	48	10			bcz8	Cum.Sti. Nimb.	1	
10															
11													Sti.Cu.		
Noon.			"	"	32^{2}		30.13	44	10			acz8	Nimb.	0	

Heuristic Geometric Information



				WINDS.			BAROMETER.		TEMPERATURE.						
Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	Direction. Magnetic	Miles ~~Force~~. during Interval.	Leeway.	Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27^{3}		30.22	50	8	.		bo	Ci:Cum Cum:St	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			ao	St. Nimb	0	
7															
8															
9			"	"	21		30.17	48	10			bczg	St. Cum:St. Nimb	1	
10															
11															
Noon.			"	"	32^{2}		30.13	44	10			aczg	St:Cum Nimb	0	

Log-linear model



Textlines with HTR transcription

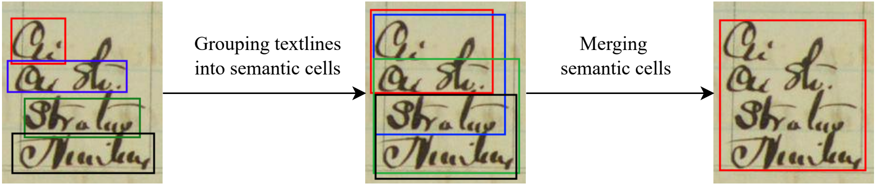
Grouping textlines

Replacing quotation marks

Performing information extraction

Database

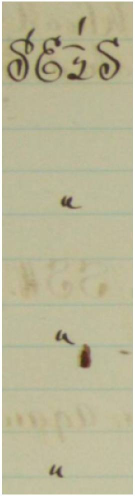
Log-linear model: grouping textlines



Grouping textlines into semantic cells

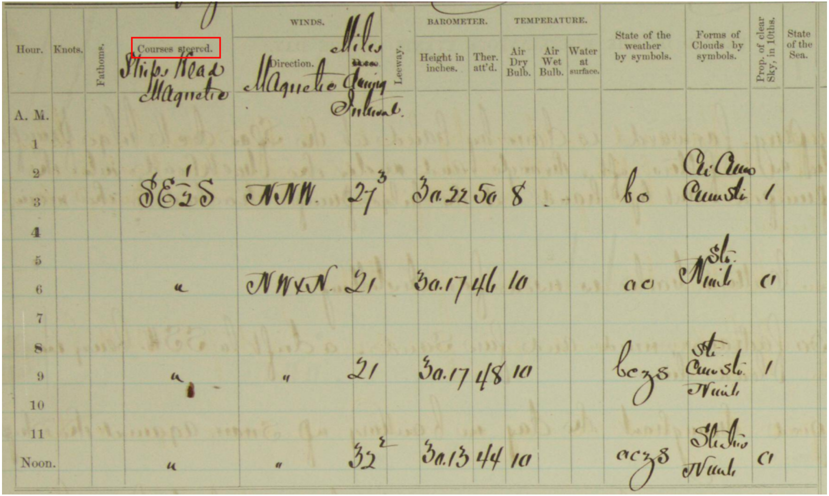
Merging semantic cells

Log-linear model: replacing quotation marks



Substitute the quotation marks by the content of the precedent cell

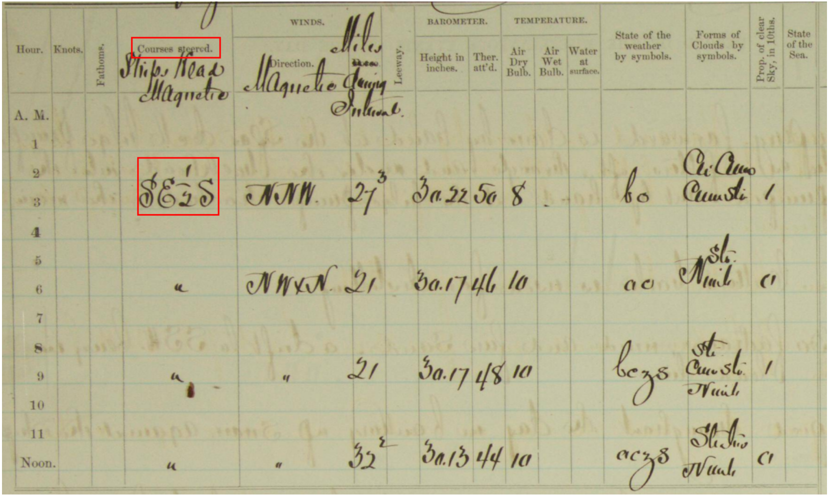
Log-linear model: performing IE



Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction. Magnetic	WINDS. Miles ~~Force.~~ Going Interval.	Leeway.	BAROMETER. Height in inches.	BAROMETER. Ther. att'd.	TEMPERATURE. Air Dry Bulb.	TEMPERATURE. Air Wet Bulb.	TEMPERATURE. Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27^{3}		30.22	50	8			bo	Ci: Cum Cum. St.	1	
4															
5															
6			“	NWxN.	21		30.17	46	10			ac	St. Nimb.	01	
7															
8															
9			“	“	21		30.17	48	10			bczs	St. Cum. St. Nimb.	1	
10															
11															
Noon.			“	“	32^{2}		30.13	44	10			aczs	St. Cirr. Nimb.	01	

Probability of being column header

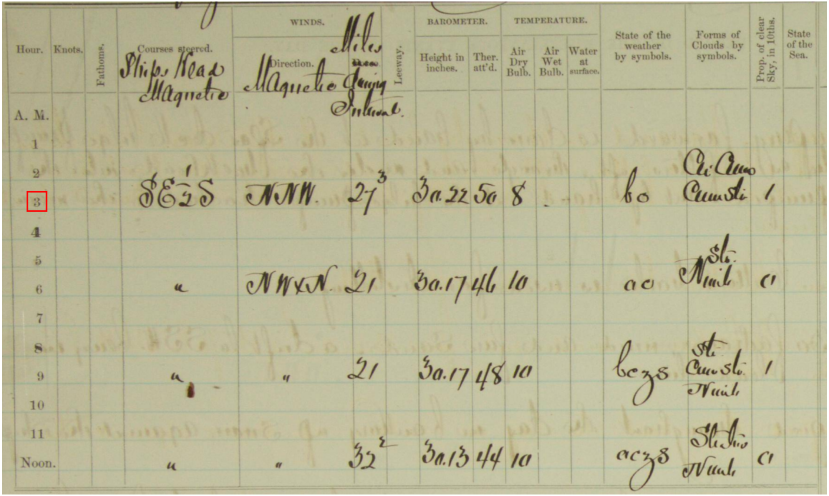
Log-linear model: performing IE



Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction. Magnetic	Miles during Interval.	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2													Ci Cum		
3			SE½S	NNW.	27^{3}		30.22	50	8	.		bo	Cum St	1	
4															
5													St		
6			"	NWxN.	21		30.17	46	10			ac	Nim	0	
7															
8													St		
9			"	"	21		30.17	48	10			bczs	Cum St	1	
10													Nimb		
11													St Ci		
Noon.			"	"	32^{2}		30.13	44	10			aczs	Nimb	0	

Probability of being aligned vertically

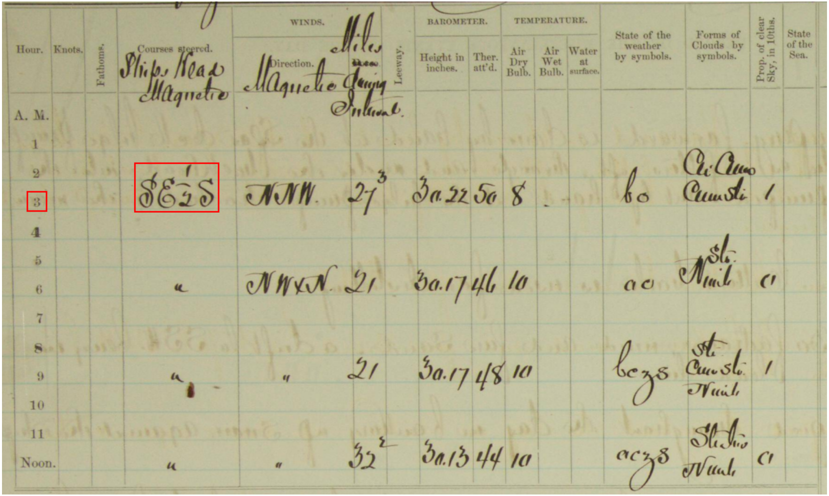
Log-linear model: performing IE



Hour.	Knots.	Fathoms.	Courses steered.	WINDS.		Leeway.	BAROMETER.		TEMPERATURE.			State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
			Ship's Head Magnetic	Direction. Magnetic	Miles ~~Force.~~ during Interval.		Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.				
A. M.															
1															
2															
3			SE½S	NNW.	27^{3}		30.22	50	8	.		bo	Ci: Cum Cum:St.	1	
4															
5															
6			"	NW×N.	21		30.17	46	10			ac	St. Nim.	0	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum St. Nim.	1	
10															
11															
Noon.			"	"	32^{2}		30.13	44	10			aczs	St. Cir. Nim.	0	

Probability of being a row header

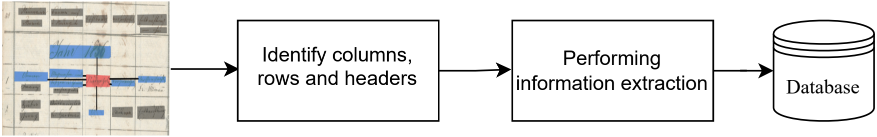
Log-linear model: performing IE



Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction. Magnetic	Miles ~~Force.~~ During Interval.	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW.	27^{3}		30.22	50	8	.		bc	Ci: Cum Cum St:	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			oc	St. Nimb	0	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum St. Nimb	1	
10															
11															
Noon.			"	"	32^{2}		30.13	44	10			oczs	St. Ci Nimb	0	

Probability of being aligned horizontally

Graph Neural Networks



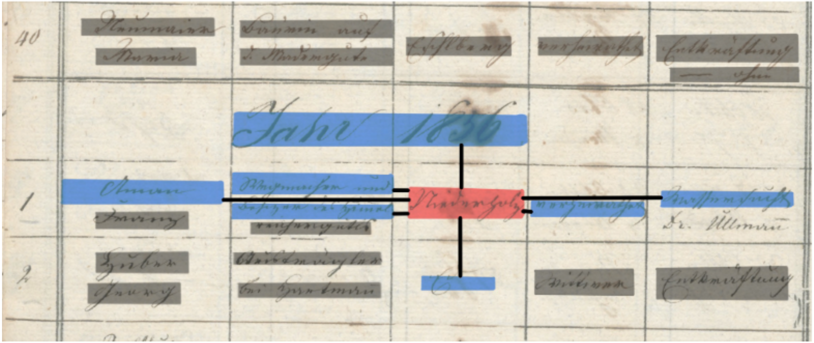
Graph of textlines

Identify columns, rows and headers

Performing information extraction

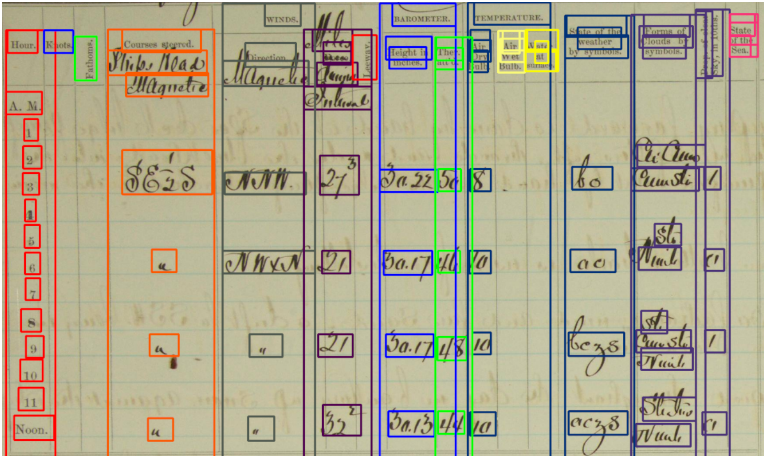
Database

Graph Neural Networks



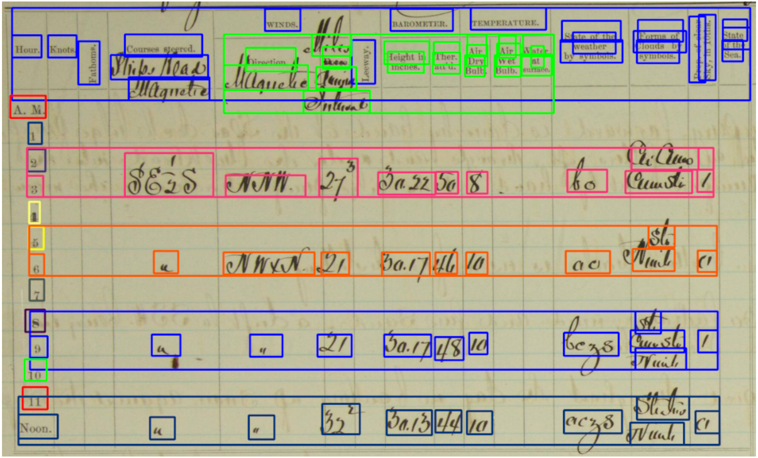
Created a graph by connecting the textlines by their line of sight

Graph Neural Networks



Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	WINDS. Direction Magnetic	Miles [illegible] Interval	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Proportion of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2			SE½S	NNW.	27 3		30.22	56	8			bc	Ci. Cum Cumuli	1	
3															
4															
5															
6			"	NW×N	21		30.17	46	10			oc	St. Nimb	10	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cumuli Nimb	1	
10															
11															
Noon.			"	"	32 2		30.13	44	10			oczs	Stratus Nimb	10	

Graph Neural Networks



Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	WINDS. Direction Magnetic	Force	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Proportion of clear sky, in tenths.	State of the Sea.
A. M.															
1															
2															
3			SE½S	NNW	27³		30.22	50	8			bo	Ci.Cum Cum.Str	1	
4															
5															
6			"	NWxN.	21		30.17	46	10			oc	Str. Nimb	c1	
7															
8															
9			"	"	21		30.17	48	10			bczs	St. Cum.Str Nimb	1	
10															
11															
Noon.			"	"	32²		30.13	44	10			oczs	Str.Cu Nimb	c1	

Jeannette Statistics

Table: Basic statistics of the HisClima database for the three partitions.

Number of:	Train	Validation	Test	Total
Pages	143	15	50	208
Lines	23 617	2 284	7 838	33 739
Running words	46 599	4 604	15 611	66 814
Lexicon	1 287	491	924	1 483
Character set size	76	76	76	76

Table: Results of text recognition.

Text type	Manuscript	Printed	Overall
WER	10.4%	1.8%	4.4%

Information Extraction

Table: Information extraction results. 95% confidence intervals are never larger than 0.01 when using GT lines and 0.02 when employing automatic lines.

Lines	Ground Thruth			Automatic		
Metric	P	R	F_1	P	R	F_1
Heuristic Geometric Information	0.79	0.78	0.78	0.64	0.55	0.59
Log-linear Model	0.87	0.79	0.83	0.77	**0.69**	**0.73**
Graph Neural Network	**0.88**	**0.83**	**0.85**	**0.78**	0.67	0.72
Oracle	0.89	0.88	0.89	0.79	0.72	0.76

Conclusions

- Machine learning approaches outperform the heuristic geometric information method.

- Log-linear model and graph neural networks have achieved a similar performance.

Future works

- Improve automatic line extraction and HTR.
- Repeat this experiment employing PrIx instead of the HTR transcripts.

Information Extraction from Handwritten Tables in Historical Documents

José Andrés[1], Jose Ramón Prieto[1], Emilio Granell[1], Verónica Romero[2], Joan Andreu Sánchez[1], Enrique Vidal[1]

[1]Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain

[2]Departament d'Informàtica, Universitat de València, Spain.

May 23rd, 2022

HTR performance

Table: Results of text recognition. 95% confidence intervals are never larger than 1.1% for manuscript text, 0.4% for printed text and 0.5 % overall.

Text type	Manuscript	Printed	Overall
CER	5.7%	1.5%	2.0%
WER	10.4%	1.8%	4.4%