UNIVERSITÀ DEGLI STUDI MILANO BICOCCA

Analisi e modellizzazione della variabile "Temperatura"

Ludovico Gatti

Matricola: 805841

Indice:

Dataset

Il dataset contiene **9358** osservazioni di sensori chimici incorporati in un dispositivo multisensore della qualità dell'aria. Il dispositivo si trovava sul campo in un'area significativamente inquinata, a livello stradale, all'interno di una città italiana. I dati sono stati registrati da marzo 2004 a febbraio 2005 (un anno). Il dataset è stato raccolto da Saverio De Vito (saverio.devito@enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development.

La fonte dei dati è:

https://archive.ics.uci.edu/ml/datasets/Air+quality

Variabili:

- **Date**: data (in formato DD/MM/YYYY)

- **Time:** (in formato HH.MM.SS)

- **CO(GT):** Concentrazione oraria media effettiva di CO in mg/m^3

- **PT08.S1(CO):** risposta oraria media del sensore (ossido di stagno) (con

target CO)

- NMHC(GT): Concentrazione media oraria di idrocarburi non Metanici in microg/m^3

- C6H6(GT): Concentrazione oraria media di benzene in microg/m^3 (reference analyzer)

- PT08.S2(NMHC): (titanio) risposta oraria media (NMHC target nominale)

- NOx(GT): Concentrazione media oraria di NOx in ppb.

- PT08.S3(NOx): (ossido di tungsteno) risposta oraria media del sensore (NOx terget nominale)

- NO2(GT): Concentrazione di NO2 media oraria in microg/m^3

- PT08.S4(NO2): (ossido di tungsteno) risposta oraria media del sensore (NO2 terget nominale)

- PT08.S5(O3): (ossido di indio) risposta media oraria del sensore (O3 target nominale)

- T: Temperature in °C

- RH: umidità relativa (%)

- AH: umidità assoluta (%)

Trasformazioni sul dataset:

Nelle informazioni generali sul dataset era specificato che i valori missing sono stati contrassegnati con il valore -200. Per evitare influenze di tali valori sul modello di regressione ho provveduto a eliminare tali osservazioni. Portando cosi il dataset da 9357 a **828** osservazioni "utilizzabili".

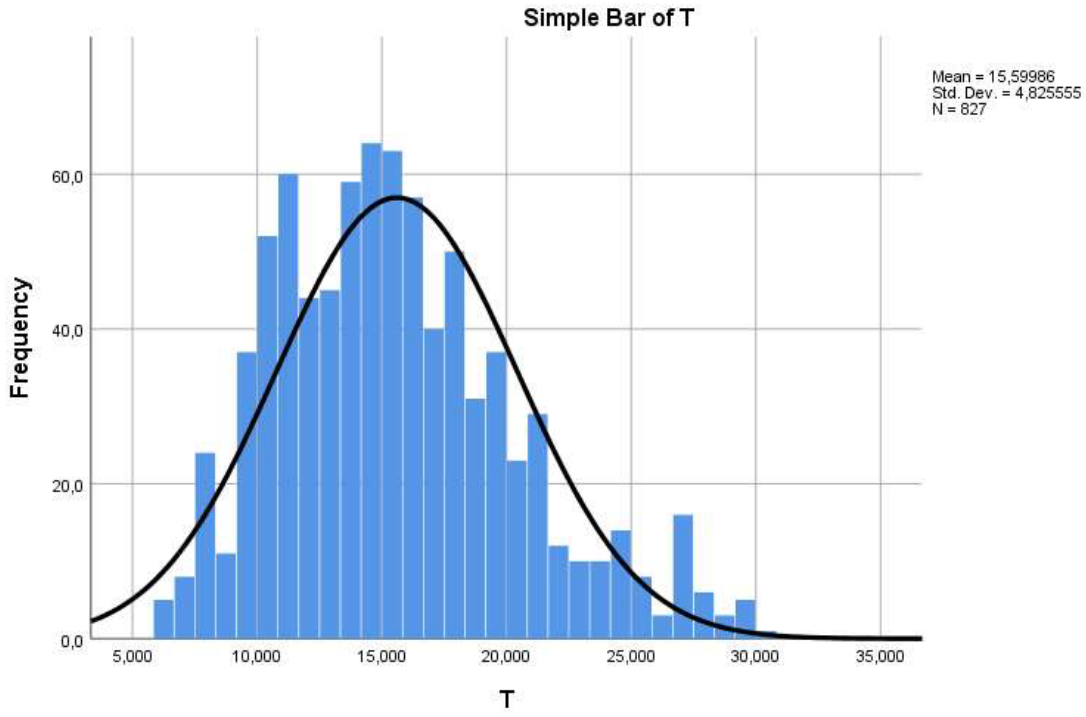
Ho preferito eliminare dell'ingagine la variabile data e la variabile time, trattandosi di un modello di regressione.

Statistiche sulla variabile Target:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
T	827	6,275	30,0	15,5	4,825	23,28	,625	,085	,075	,170
Valid N	827									

La variabile target presenta una media di 15,5 gradi con una deviazione standard di circa 5 gradi. Presenta una leggera asimmetria positiva e una minima kurtosi che ne fanno assumere l'andamento quasi campanulare come si può evincere dal grafico stottostante.



Modelli

L’intento dell’analisi è costruire un **modello di regressione multipla** capace di spiegare la dipendenza dell’andameto della temperatura dagli atri fattori inquinanti.

Verrà usato il software **IBM SPSS.**

Si procede nel costruire il primo modello di regressione (modello completo) con tutte le variabili esplicative descritte precedentemente.

Risultati del primo modello (modello completo)

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	AH, NO2(GT), RH, NMHC(GT), PT08.S3(NOx), PT08.S5(O3), NOx(GT), PT08.S1(CO), C6H6(GT), CO(GT), PT08.S4(NO2), PT08.S2(NMHC)[b]	.	Enter

a. Dependent Variable: T

b. All requested variables entered.

Dal testing si può evincere che tutte le variabili selezionate sono entrate a far parte del modello.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,979[a]	,958	,957	,994823

a. Predictors: (Constant), AH, NO2(GT), RH, NMHC(GT), PT08.S3(NOx), PT08.S5(O3), NOx(GT), PT08.S1(CO), C6H6(GT), CO(GT), PT08.S4(NO2), PT08.S2(NMHC)

b. Dependent Variable: T

Si nota un **R-Square** molto alto: 95,8%. Il modello riesce a spiegare quasi tutta la variabilità della variabile **T** (temperatura).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18428,624	12	1535,719	1551,744	,000[b]
	Residual	805,594	814	,990		
	Total	19234,218	826			

a. Dependent Variable: T

b. Predictors: (Constant), AH, NO2(GT), RH, NMHC(GT), PT08.S3(NOx), PT08.S5(O3), NOx(GT), PT08.S1(CO), C6H6(GT), CO(GT), PT08.S4(NO2), PT08.S2(NMHC)

Analizzando la varianza con l'analisi *Anova*, si nota che la statistica **F di Fisher** assume il valore di 1551,744, e il relativo ***p-value*** risulta basso. Quindi, il modello è significativo per un alpha di 0,001.

Possiamo quindi escludere l'ipotesi nulla, per la quale i coefficienti di regressione siano tutti nulli.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	9,601	1,784		5,382	,000
	CO(GT)	-,879	,152	-,257	-5,789	,000
	PT08.S1(CO)	-,001	,001	-,051	-1,512	,131
	NMHC(GT)	-,001	,000	-,024	-1,260	,208
	C6H6(GT)	,032	,058	,049	,551	,582
	PT08.S2(NMHC)	,010	,002	,578	5,913	,000
	NOx(GT)	,005	,002	,078	2,739	,006
	PT08.S3(NOx)	,003	,001	,171	5,358	,000
	NO2(GT)	,009	,003	,059	2,977	,003
	PT08.S4(NO2)	-,003	,001	-,195	-3,233	,001
	PT08.S5(O3)	-,001	,000	-,099	-4,097	,000
	RH	-,327	,004	-1,034	-87,274	,000
	AH	20,020	,423	,741	47,361	,000

a. Dependent Variable: T

Dalla tabella che riporta i coefficienti del modello di regressione si può osservare la colonna dei test di significativita degli stessi. Si evince che alcuni coeffcienti presentano un livello di significatività superiore allo 0,05%.

Provvedo ad eliminare le variabili che possiedono tali coefficienti e a calcolare un nuovo modello di regressione (modello ristretto).

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5,25290	27,53063	15,59986	4,723419	827
Residual	-2,533635	5,483279	,000000	,987570	827
Std. Predicted Value	-2,191	2,526	,000	1,000	827
Std. Residual	-2,547	5,512	,000	,993	827

a. Dependent Variable: T

Risultati modello ristretto

Dal modello completo sono state eliminate le variabili: **PT08.S1(CO), NMHC(GT), C6H6(GT).**

Le variabili inserite nel modello si possono osservare nel primo output.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	AH, NO2(GT), RH, CO(GT), PT08.S3(NOx), PT08.S5(O3), NOx(GT), PT08.S4(NO2), PT08.S2(NMHC)[b]	.	Enter

a. Dependent Variable: T

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,979[a]	,958	,957	,995158

a. Predictors: (Constant), AH, NO2(GT), RH, CO(GT), PT08.S3(NOx), PT08.S5(O3), NOx(GT), PT08.S4(NO2), PT08.S2(NMHC)

b. Dependent Variable: T

Come si può osservare il coefficiente R-Square non è peggiorato dopo l'eliminazione delle variabili.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18425,110	9	2047,234	2067,203	,000[b]
	Residual	809,108	817	,990		
	Total	19234,218	826			

a. Dependent Variable: T

b. Predictors: (Constant), AH, NO2(GT), RH, CO(GT), PT08.S3(NOx), PT08.S5(O3), NOx(GT), PT08.S4(NO2), PT08.S2(NMHC)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8,772	1,009		8,697	,000
	CO(GT)	-,942	,121	-,275	-7,764	,000
	PT08.S2(NMHC)	,011	,001	,599	8,928	,000
	NOx(GT)	,005	,002	,079	2,800	,005
	PT08.S3(NOx)	,003	,000	,173	7,483	,000
	NO2(GT)	,008	,003	,051	2,789	,005
	PT08.S4(NO2)	-,003	,001	-,198	-3,629	,000
	PT08.S5(O3)	-,001	,000	-,111	-5,245	,000
	RH	-,327	,004	-1,035	-90,506	,000
	AH	19,934	,420	,737	47,423	,000

a. Dependent Variable: T

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5,25238	27,62047	15,59986	4,722968	827
Residual	-2,527952	5,494860	,000000	,989722	827
Std. Predicted Value	-2,191	2,545	,000	1,000	827
Std. Residual	-2,540	5,522	,000	,995	827

a. Dependent Variable: T

Scelta del modello

Si vuole verificare se il modello intero non sia significativamente migliore di quello ristretto. Si procederà ad eseguire **il test F** la cui statistica test è:

$$F = \frac{\frac{[\text{DevRes}]_r - [\text{DevRes}]_i}{Vi - Vr}}{\frac{[DevRes]_i}{(n - Vi)}}$$

dove **n** è la dimensione del campione, **Vi** è il numero di variabili esplicative nel modello completo e **Vr** è il numero di parametri nel modello ristretto; che si distribuisce sotto l'ipotesi nulla come una F di Fisher con **(Vi-Vr , n – Vi)** gradi di libertà.

La statistica ha un valore di

$$F = \frac{\frac{809{,}108 - 805{,}594}{12-9}}{\frac{805{,}594}{(828-12)}} = 0{,}000903$$

Tale valore è minore del valore critico che risulta essere **8.53** per un F-Fisher con un livello di significatività alpha pari a 0.05 con (3 , 816) gradi di libertà. Risulta quindi accettata l'ipotesi nulla: Il modello ristretto risulta migliore.

Conclusioni sul modello:

Analizziamo i coefficienti del modello ristretto per trarre le conclusioni finali.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8,772	1,009		8,697	,000
	CO(GT)	-,942	,121	-,275	-7,764	,000
	PT08.S2(NMHC)	,011	,001	,599	8,928	,000
	NOx(GT)	,005	,002	,079	2,800	,005
	PT08.S3(NOx)	,003	,000	,173	7,483	,000
	NO2(GT)	,008	,003	,051	2,789	,005
	PT08.S4(NO2)	-,003	,001	-,198	-3,629	,000
	PT08.S5(O3)	-,001	,000	-,111	-5,245	,000
	RH	-,327	,004	-1,035	-90,506	,000
	AH	19,934	,420	,737	47,423	,000

a. Dependent Variable: T

Il coefficiente della variabile che indica la concentrazione di **CO** è negativo. Segno che all'aumentare della CO nell'aria, la temperatura tende a diminuire. Negativi sono anche i coefficienti dell'umidità relativa (RH), monossido di azoto (PT08.S4(NO2)) e dell'ozono (PT08.S5(O3)). Quindi

all’aumentare di queste ultime, la temperatura diminuirebbe (con intensità più o meno modeste. Meno di mezzo grado nel caso dell’umidità relativa.)

Mentre i coefficienti delle restanti variabili (**PT08.S2(NMHC), NOx(GT), PT08.S3(NOx), NO2(GT)** e **AH**) sono positivi. Segno che il loro incremento tende a far aumentare il valore della variabile target. Il valore dei coefficienti risulta essere modesto eccezione fatta per la variabile **AH** (umidità assoluta) il cui valore (19,9) sembra essere relativamente alto. Segno che l’umidità ha molta influenza sull’andamento della temperatura.

Analisi dei residui:

Si vuole inoltre studiare l'andamento dei residui generati del modello scelto (modello ristretto).

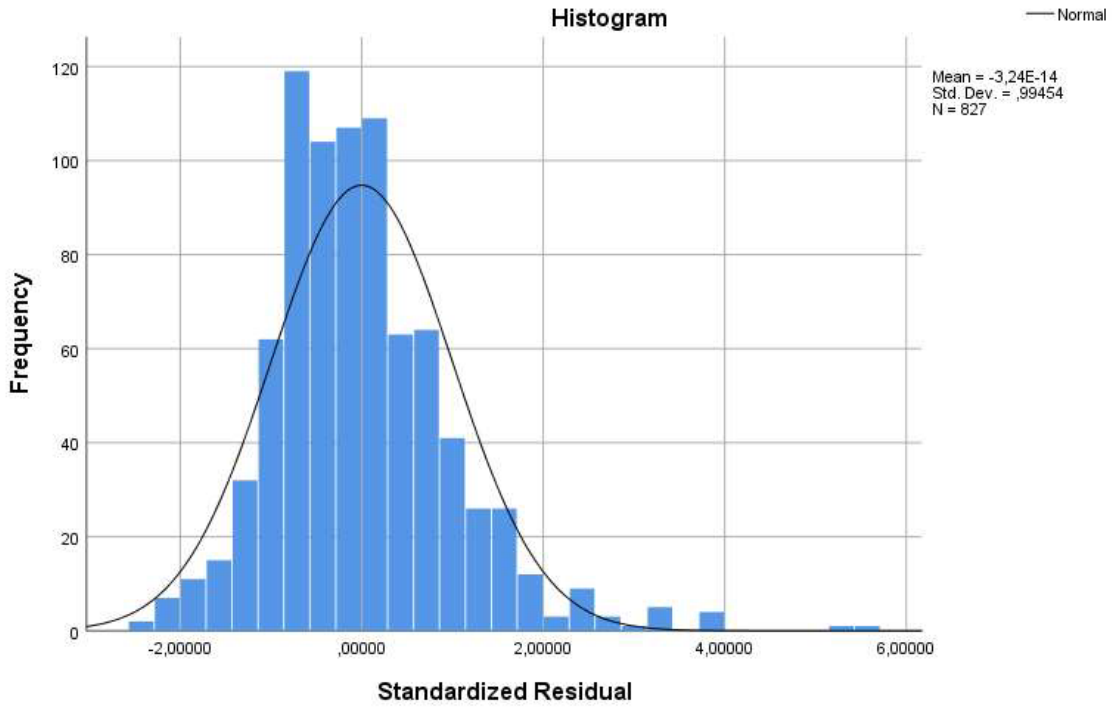
Case Processing Summary

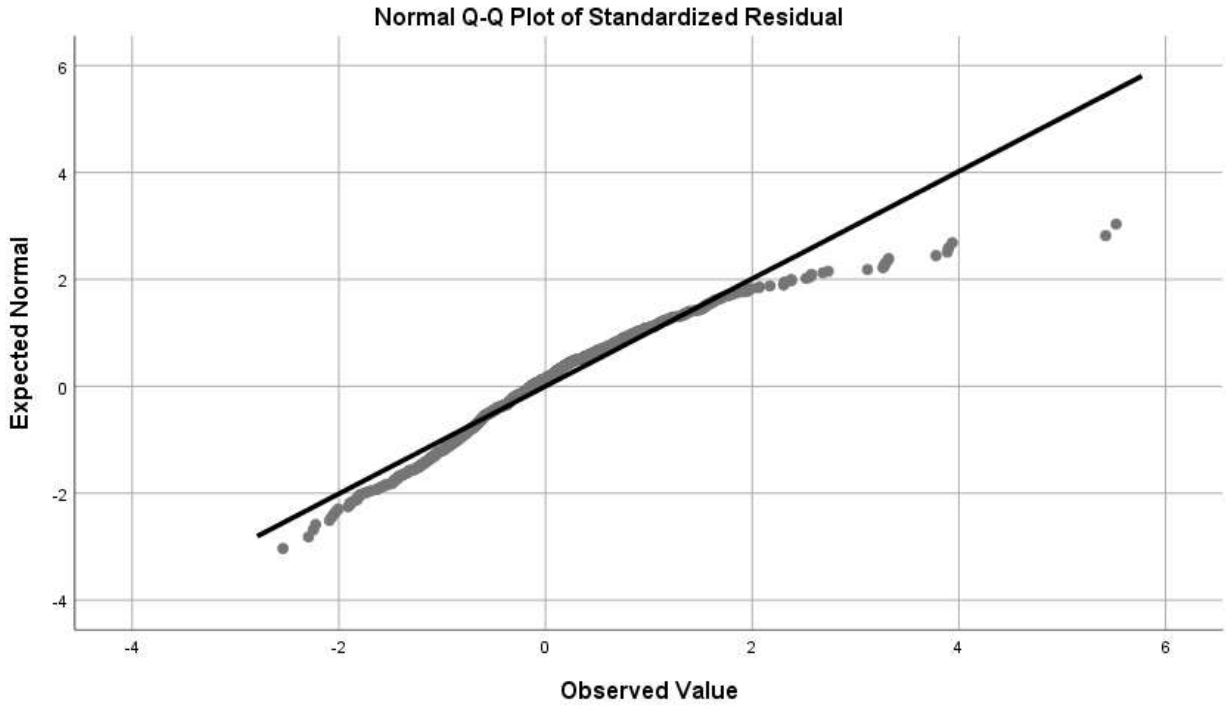
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Standardized Residual	827	87,9%	114	12,1%	941	100,0%

Descriptives

			Statistic	Std. Error
Standardized Residual	Mean		,0000000	,03458344
	95% Confidence Interval for Mean	Lower Bound	-,0678818	
		Upper Bound	,0678818	
	5% Trimmed Mean		-,0493973	
	Median		-,1349517	
	Variance		,989	
	Std. Deviation		,99453714	
	Minimum		-2,54025	
	Maximum		5,52159	
	Range		8,06184	
	Interquartile Range		1,16275	
	Skewness		1,102	,085
	Kurtosis		3,138	,170

Come si evince i residui hanno media zero. L'andamento delle frequenze possiede un'asimmetria positiva e una Kurtosi abbastanza pronunciata.





Il grafico dei quantili ci mostra che essi si distribuiscono più o meno intorno alla retta.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,082	827	,000	,946	827	,000

a. Lilliefors Significance Correction

Il test di normalità di Shapiro-Wilk mostra una valore molto vicino ad 1. Tuttavia presenta un un livello di significatività p-value molto basso (molto meno del 5%) e ciò ci porterebbe ad escludere l'ipotesi nulla: i dati sono significativamente diversi dalla distribuzione normale.

Campione di osservazioni:

Si riporta di seguito un campione di 100 osservazioni utilizzate per l'analisi.

CO (GT)	PT08.S1 (CO)	NMHC (GT)	C6H6 (GT)	PT08.S2 (NMHC)	NOx(GT)	PT08.S3 (NOx)	NO2 (GT)	PT08.S4 (NO2)	PT08.S5 (O3)	T	RH	AH
2,6	1360	150	11,9	1046	166	1056	113	1692	1268	13,6	48,9	0,7578
2	1292	112	9,4	955	103	1174	92	1559	972	13,3	47,7	0,7255
2,2	1402	88	9,0	939	131	1140	114	1555	1074	11,9	54,0	0,7502
2,2	1376	80	9,2	948	172	1092	122	1584	1203	11,0	60,0	0,7867
1,6	1272	51	6,5	836	131	1205	116	1490	1110	11,2	59,6	0,7888
1,2	1197	38	4,7	750	89	1337	96	1393	949	11,2	59,2	0,7848
1,2	1185	31	3,6	690	62	1462	77	1333	733	11,3	56,8	0,7603
1	1136	31	3,3	672	62	1453	76	1333	730	10,7	60,0	0,7702
0,9	1094	24	2,3	609	45	1579	60	1276	620	10,7	59,7	0,7648
0,7	1066	8	1,1	512	16	1918	28	1182	422	11,0	56,2	0,7366
0,7	1052	16	1,6	553	34	1738	48	1221	472	10,5	58,1	0,7353
1,1	1144	29	3,2	667	98	1490	82	1339	730	10,2	59,6	0,7417
2	1333	64	8,0	900	174	1136	112	1517	1102	10,8	57,4	0,7408
2,2	1351	87	9,5	960	129	1079	101	1583	1028	10,5	60,6	0,7691
1,7	1233	77	6,3	827	112	1218	98	1446	860	10,8	58,4	0,7552
1,5	1179	43	5,0	762	95	1328	92	1362	671	10,5	57,9	0,7352
1,6	1236	61	5,2	774	104	1301	95	1401	664	9,5	66,8	0,7951
1,9	1286	63	7,3	869	146	1162	112	1537	799	8,3	76,4	0,8393
2,9	1371	164	11,5	1034	207	983	128	1730	1037	8,0	81,1	0,8736
2,2	1310	79	8,8	933	184	1082	126	1647	946	8,3	79,8	0,8778
2,2	1292	95	8,3	912	193	1103	131	1591	957	9,7	71,2	0,8569
2,9	1383	150	11,2	1020	243	1008	135	1719	1104	9,8	67,6	0,8185
4,8	1581	307	20,8	1319	281	799	151	2083	1409	10,3	64,2	0,8065
6,9	1776	461	27,4	1488	383	702	172	2333	1704	9,7	69,3	0,8319
6,1	1640	401	24,0	1404	351	743	165	2191	1654	9,6	67,8	0,8133
3,9	1313	197	12,8	1076	240	957	136	1707	1285	9,1	64,0	0,7419
1,5	965	61	4,7	749	94	1325	85	1333	821	8,2	63,4	0,6905
1	913	26	2,6	629	47	1565	53	1252	552	8,2	60,8	0,6657
1,7	1080	55	5,9	805	122	1254	97	1375	816	8,3	58,5	0,6438
1,9	1044	53	6,4	829	133	1247	110	1378	832	7,7	59,7	0,6308
1,4	988	40	4,1	718	82	1396	91	1304	692	7,1	61,8	0,6276
0,6	847	7	1,0	501	30	1895	44	1155	394	6,3	65,0	0,6233
0,8	927	17	1,8	571	56	1685	71	1223	487	6,8	62,9	0,6234
1,4	1091	33	4,4	730	109	1387	104	1361	748	6,4	65,1	0,6316
4,4	1587	202	17,9	1236	307	897	141	1900	1400	7,3	63,1	0,6499
3,1	1350	208	14,0	1118	187	912	122	1712	1237	13,2	41,7	0,6320
2,7	1263	166	11,6	1037	216	969	143	1598	1167	14,3	38,4	0,6243
2,1	1206	114	10,2	986	143	1035	113	1537	959	15,0	36,5	0,6195
2,5	1252	140	11,0	1016	160	1008	116	1593	983	16,1	34,5	0,6262
2,7	1287	169	12,8	1078	163	949	123	1660	1061	16,3	35,7	0,6560
2,9	1353	185	14,2	1122	190	922	126	1740	1139	15,8	37,0	0,6610
2,8	1309	165	12,7	1073	178	954	120	1657	1112	15,9	37,2	0,6657
2,4	1274	133	11,7	1041	150	1006	119	1610	994	16,9	34,3	0,6549
3,9	1510	233	19,3	1277	206	812	149	1910	1410	15,1	39,6	0,6766
3,7	1525	242	18,2	1246	202	821	145	1847	1448	14,4	43,4	0,7084
6,6	1843	488	32,6	1610	340	624	170	2390	1887	12,9	50,5	0,7478
4,4	1598	333	20,1	1299	274	752	149	1941	1627	12,1	53,3	0,7536

3,5	1484	215	14,3	1127	253	839	139	1723	1491	11,0	59,1	0,7740
5,4	1677	367	21,8	1346	300	741	134	2062	1657	9,7	64,6	0,7771
2,7	1280	122	9,6	964	193	963	113	1544	1285	9,5	64,1	0,7597
1,9	1196	67	7,4	873	139	1071	97	1463	1144	9,1	63,9	0,7423
1,6	1184	43	5,4	782	83	1176	82	1365	1043	8,8	63,9	0,7256
1	978	30	2,6	625	62	1420	65	1274	819	8,3	63,6	0,6982
1,2	1100	27	2,9	646	53	1406	60	1268	835	7,2	67,5	0,6887
1,5	1112	47	5,1	770	139	1228	77	1409	940	6,3	71,9	0,6932
2,7	1336	132	11,8	1043	256	935	96	1678	1192	6,5	71,6	0,6945
3,7	1408	239	15,1	1153	295	830	119	1777	1411	9,6	59,7	0,7124
3,2	1447	160	12,9	1081	250	869	126	1667	1465	12,4	51,2	0,7335
4,1	1542	283	16,1	1184	296	808	158	1780	1583	15,6	42,2	0,7451
3,6	1451	210	14,0	1117	239	875	161	1679	1387	18,4	33,8	0,7090
2,8	1328	154	12,3	1059	153	987	124	1600	1101	19,4	31,3	0,6950
2	1207	112	8,6	924	118	1088	102	1488	850	18,0	34,8	0,7127
2	1240	108	9,2	947	119	1049	116	1532	947	18,4	33,6	0,7042
2,5	1306	111	10,2	987	138	1004	124	1554	1078	17,6	35,1	0,7012
2,3	1326	97	10,6	1000	148	976	125	1602	1084	16,7	37,8	0,7117
3,2	1473	191	15,5	1163	227	831	148	1779	1395	16,1	41,0	0,7451
4,2	1609	258	19,6	1286	277	758	165	1922	1612	15,8	42,4	0,7569
4,2	1611	284	19,2	1274	279	754	161	1915	1697	15,7	44,1	0,7786
4,2	1621	269	18,3	1247	283	762	159	1860	1886	15,3	46,8	0,8091
3,1	1444	180	13,1	1089	214	844	143	1748	1624	14,6	48,6	0,8060
2,6	1418	116	10,9	1010	172	892	130	1603	1536	14,7	49,3	0,8193
2,9	1534	93	11,0	1013	190	889	129	1611	1535	13,9	53,6	0,8498
2,8	1484	131	11,9	1045	174	880	119	1624	1530	14,6	51,5	0,8536
2,5	1367	92	8,6	925	128	953	104	1543	1337	12,5	58,9	0,8537
1,2	1062	32	3,7	691	53	1272	70	1377	929	11,5	63,1	0,8533
1	1076	29	2,5	618	44	1395	63	1333	872	11,6	62,2	0,8473
0,9	1028	27	2,4	615	74	1384	67	1340	853	10,4	67,6	0,8530
1,4	1155	36	4,2	722	101	1225	84	1414	959	11,6	62,7	0,8530
1,6	1235	57	6,4	828	118	1055	83	1527	1093	12,4	60,0	0,8627
2,2	1332	129	8,6	923	144	952	98	1614	1225	14,5	53,1	0,8728
2,8	1445	148	10,9	1009	176	878	114	1696	1355	16,9	46,1	0,8789
2,8	1416	145	10,7	1002	161	907	119	1677	1262	19,3	38,3	0,8474
2	1281	93	7,5	880	113	1084	104	1525	980	21,2	31,4	0,7812
1,8	1207	84	7,5	879	103	1104	102	1490	872	21,4	30,2	0,7616
1,9	1258	99	8,2	906	112	1081	107	1511	900	21,9	29,0	0,7525
3	1458	150	11,9	1045	170	974	129	1646	1099	22,2	28,4	0,7516
2,9	1438	156	12,0	1051	180	943	128	1668	1206	21,3	30,8	0,7696
2,5	1478	122	12,2	1055	160	929	121	1671	1262	19,7	36,7	0,8307
4,6	1808	262	20,6	1312	261	753	157	1993	1698	18,4	41,7	0,8732
5,9	1898	341	23,1	1381	325	681	173	2103	1905	17,6	46,1	0,9210
3,4	1560	214	14,7	1140	217	784	146	1818	1648	16,7	49,6	0,9320
2,1	1324	100	9,0	940	146	924	121	1587	1423	16,3	51,0	0,9341
2,2	1349	79	8,8	933	152	933	119	1617	1349	14,7	55,9	0,9314
1,8	1239	66	7,4	872	104	985	99	1547	1250	14,8	54,7	0,9164
1,8	1239	73	6,9	853	106	1010	93	1543	1174	14,0	57,0	0,9094
1,8	1224	66	7,0	855	108	998	88	1566	1149	13,4	61,3	0,9361
1	1075	39	3,9	703	88	1156	74	1464	1010	11,9	67,4	0,9375
1,4	1157	51	6,4	830	138	1030	80	1584	1083	11,4	70,5	0,9475
2,2	1314	107	9,7	966	228	897	89	1710	1235	11,3	70,2	0,9401