Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Petr Jelínek

Ověření normality rozdělení některých klimatických prvků

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Radan Huth, DrSc.

Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 5.8.2008 Petr Jelínek

Contents

Název práce: Ověření normality rozdělení některých klimatických prvků
Autor: Petr Jelínek
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Radan Huth, DrSc., Ústav fyziky atmosféry
E-mail vedoucího: huth@ufa.cas.cz
Abstrakt: V klimatologii se běžně předpokládá, že teplota má normální rozdělení. Cílem studie provedené na datech z celé Evropy za 20. století je ověřit, zda je taková aproximace vhodná, případně navrhnout lepší model a zjistit, zda se rozdělení teplot liší v jednotlivých regionech. V kapitole 2 je popsána příprava dat. Kapitola 3 obsahuje výsledky testů normality založených na momentech vyššího řádu a testů dobré shody založených na empirické distribuční funkci. V kapitole 4 je představeno složené normální rozdělení jako kandidát na lepší model. Tato hypotéza je testována Engelmanovým–Hartiganovým testem a iterativním EM algoritmem. Kapitola 5 shrnuje dosažené výsledky.
Klíčová slova: Normalita, teplota, Evropa, rozdělení

Title: Testing for Normality of Certain Climatology Quantities
Author: Petr Jelínek
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Radan Huth, DrSc., Institute of Atmospheric Physics
Supervisor's e-mail address: huth@ufa.cas.cz
Abstract: Normality of daily temperature is often presumed in climatology. This study aims to verify the adequacy of such a model and possibly design a better one. A set of 20^{th} century temperature data from all parts of Europe is examined. An effort is made to find out, whether regional differences in temperature distribution exist. Chapter 2 describes the preliminary trasformation of the data. In Chapter 3, several normality tests are conducted, based on both higher order statistical moments and EDF goodness-of-fit. Chapter 4 introduces finite normal mixtures, Engelman–Hartigan test and an iterative EM algorithm are applied. Chapter 5 summarizes the results.
Keywords: Normality, temperature, Europe, distribution

Chapter 1

Introduction

The objective of this work is to examine statistical behaviour of daily temperatures. Temperature is one of the fundamental quantities used in climatology and a statistical analysis of a large sample is often used. Unfortunately, statistical methods usually rely on premises that might not be valid for climatological data. The most obvious example is presumed normality in temperatures. While the temperatures follow the common pattern of rare extreme values, the distribution is often skewed or multimodal. As the most common and easily accessible statistical methods often require normally distributed data and many are far from robust, the non-normal input can cause huge errors in results, leading to ill-advised conclusions. It is therefore useful to examine if there are any circumstances where normality can be assumed.

It would be foolish to expect obtaining a T-shirt formula to cover temperature distribution accross the globe and time of the year. In fact, this sounds even less likely than confirming that assuming normality for any time and place is adequate. This work has two realistic goals.

- Find out if normality can be safely assumed at certain regions and/or situations (based on temperature type and season).
- Identify patterns that could be relevant for temperature behaviour.[1] It is unlikely to get firm results just based on statistical analysis but such

[1] Such as well-known skewness in minimum temperatures in cold areas.

patterns should be a precious food for thought for an expert climatologist who could be able to explain the issue based on climatological reasoning.

Several studies adressed the issue of normality in temperatures. Most notably, Harmel et al. [6] compared actual data from 15 U.S. sites over 30 years with simulations by weather generators. Actual data were found to be skewed in a way that was not reproduced by the simulation. Number of cold events (especially in winter) tends to be underestimated by generated values and the generated maximum temperatures were significantly higher than observed records. Similar study oriented at higher order statistical moments was perfomed by Huth et al. [7] on 6 stations in central Europe giving comparisons between different methods of modelling temperatures. Gong and Ho [5] conducted a study on 155 Chinese and Korean stations over the winter months to see that the recent warming trend does not only change mean numbers but also influences several other factors and leads to a notable decrease in intra-seasonal variability of temperatures, leading to a more stable weather over East Asia. From a theoretical standpoint, Toth and Szentimrey [12] developed so-called binormal distribution[2] and suggested to use it instead of normal distribution when asymmetry needs to be considered.

[2]Binormal distribution is NOT the same as two-component normal mixture distribution that is discussed later in this paper.

Chapter 2

Data Sets

2.1 Origin of Data

This project uses temperature measurements from 98 European stations that cover all the European mainland (the only notable exception is a relatively narrow stripe starting in Poland and going east through Belarus and Ukraine), Iceland, Ireland and three stations located in the ocean or on small islands [8]. There are three values measured at each station - minimum temperature, mean temperature and maximum temperature. Maximum time frame is between January 1, 1901 and December 31, 1999 but most series start and/or end on different days and there are also missing values.

Table 2.1 shows the list of stations together with their geographical location and sample sizes for different seasons and temperature type. Summer seasons tend to have slightly more observations as the season is longer. Maximum and minimum temperatures usually have about the same number of observations, although there are exceptions.[1]

All computation was performed by R software [9]. `fields`, `nortest` and `mclust` packages were used.

It needs to be noted that mean temperature is fundamentally different to minimum and maximum tepmerature. The latter two are a single measurement at a given time of the day. On the other hand, mean temperature is

[1] Baia-Mare and VF-Omul in Romania and Kursk, Russia

CNTRY	LOCATION	LAT	LON	HEIGHT	min-f	min-w	min-lw	mean-f	mean-w	mean-lw
AT	KREMS	48.05	14.13	383	35770	8820	14798	35770	8820	14798
BE	UCCLE	50.8	4.35	100	36134	8909	14948	0	0	0
CR	ZAGRE	45.82	15.98	157	35405	8730	14647	35405	8730	14647
CZ	PRAHA	50.09	14.42	191	36135	8910	14949	36135	8910	14949
DK	VESTE	56.77	8.32	18	35769	8820	14797	0	0	0
DK	NORDB	55.45	8.4	4	35638	8756	14703	0	0	0
DK	KODAN	55.68	12.53	9	35219	8699	14586	0	0	0
FI	HELSI	60.17	24.95	4	17885	4410	7399	17867	4404	7391
FI	JYVAS	62.4	25.68	137	17885	4410	7399	17885	4410	7399
FI	SODAN	67.37	26.65	179	33345	8204	13761	0	0	0
FR	MARSE	43.31	5.4	75	36129	8907	14944	0	0	0
FR	BOURG	47.07	2.37	161	31755	7830	13137	0	0	0
FR	TOULO	43.62	1.38	152	33074	8188	13733	0	0	0
FR	BORDE	44.83	-0.7	49	29951	7379	12405	0	0	0
FR	CHATE	46.87	1.72	160	36134	8910	14948	0	0	0
FR	PERPI	42.73	2.87	43	35328	8653	14570	0	0	0
FR	LYON_	45.73	4.94	172	28949	7141	11990	0	0	0
FR	PARIS	48.82	2.33	75	36135	8910	14949	0	0	0
DE	HAMBU	53.55	9.97	26	36135	8910	14949	0	0	0
DE	BREME	53.05	8.78	4	35829	8879	14857	35798	8848	14826
DE	TRIER	49.75	6.65	144	29049	7141	11990	29049	7141	11990
DE	KARLS	49.02	8.38	114	35801	8820	14798	35801	8820	14798
DE	STUTT	48.72	9.22	401	36074	8879	14918	36074	8879	14918
DE	SCHWE	53.65	11.38	59	34219	8460	14194	34219	8460	14194
DE	DRESD	51.12	13.68	246	30295	7470	12533	30295	7470	12533
DE	BERLI	52.45	13.3	55	35945	8910	14949	35940	8910	14944
DE	POTSD	52.38	13.07	81	36135	8910	14949	36135	8910	14949
DE	BAMBE	49.88	10.88	282	36135	8910	14949	36135	8910	14949
DE	ZUGSP	47.42	10.98	2960	36029	8910	14949	36029	8910	14949
DE	HOHEN	47.8	11.02	977	36135	8910	14949	36124	8910	14949
DE	MUNCH	48.17	11.5	515	34949	8640	14466	34980	8640	14466
DE	JENA_	50.93	11.58	155	36042	8879	14918	36040	8878	14917
GR	LARIS	39.65	22.45	74	16054	3955	6638	15690	3866	6488
GR	HELLI	37.9	23.75	15	16056	3959	6641	16057	3959	6641
GR	HERAK	35.33	25.18	39	16060	3960	6644	15938	3929	6583
IS	REYKJ	64.13	-21.9	52	17763	4379	7338	17763	4379	7338
IS	STYKK	65.08	-22.73	8	17729	4378	7336	17725	4379	7338
IS	DALAT	65.27	-13.58	9	17336	4279	7176	17671	4379	7338
IS	VESTM	63.4	-20.28	118	17701	4379	7338	17701	4379	7338
IE	VALEN	51.94	-10.24	9	21992	5431	9121	21992	5431	9121
IE	BIRR_	53.09	-7.89	70	16516	4081	6856	16516	4081	6856
IE	MALIN	55.37	-7.34	20	16305	3991	6705	16305	3991	6705
IT	VERON	45.38	10.87	68	17490	4304	7232	0	0	0
IT	ROMA_	41.78	12.58	105	17507	4320	7248	0	0	0
LT	KLAIP	55.73	21.07	6	24441	6042	10127	24344	6039	10122
LT	KAUNA	54.88	23.83	75	28913	7105	11919	29026	7157	12003
LT	VILNI	54.63	25.28	189	33320	8340	13917	33852	8494	14157
LU	LUXEM	49.62	6.22	376	19342	4768	8000	19334	4766	7993

Table 2.1: There are 98 stations examined. Last six columns show number of observations for minimum and mean temperatues in different seasons. Figures for maximum temperatures and summer seasons are similar to those for minimum temperatures and winter.

CNTRY	LOCATION	LAT	LON	HEIGHT	min-f	min-w	min-lw	mean-f	mean-w	mean-lw
NL	DEKOO	52.92	4.78	0	34037	8370	14043	34037	8370	14043
NL	DEBIL	52.1	5.18	2	36135	8910	14949	36135	8910	14949
NL	EELDE	53.13	6.58	4	34251	8401	14135	34251	8401	14135
NL	VLISS	51.45	3.6	8	34005	8370	14043	29260	7200	12080
NL	MAAST	50.92	5.78	114	34308	8460	14194	34310	8460	14194
NO	OSLO_	59.95	10.72	94	22630	5580	9362	22630	5580	9362
NO	OKSOY	58.07	8.05	9	17885	4410	7399	17885	4410	7399
NO	UTSIR	59.3	4.88	55	17885	4410	7399	17885	4410	7399
NO	GLOMF	66.82	13.98	39	15695	3870	6493	15695	3870	6493
NO	KARAS	69.47	25.52	129	15695	3870	6493	15695	3870	6493
NO	VARDO	70.37	31.08	14	17885	4410	7399	17885	4410	7399
NO	BJORN	74.52	19.02	16	20986	5189	8697	20986	5189	8697
NO	JMAYE	70.93	-8.67	10	16059	3959	6643	16059	3959	6643
PT	LISBO	38.72	-9.15	77	36133	8910	14949	0	0	0
PT	PORTO	41.13	-8.6	93	21368	5273	8849	0	0	0
PT	BEJA_	38.02	-7.87	246	15323	3780	6341	0	0	0
PT	BRAGA	41.8	-6.73	690	21389	5307	8902	0	0	0
RO	BAIAM	47.67	23.5	216	29567	7231	12202	0	0	0
RO	CLUJ_	46.78	23.57	410	23270	5670	9573	0	0	0
RO	ARAD_	46.13	21.35	117	28042	6837	11564	0	0	0
RO	VFOMU	45.45	25.45	2504	23603	5729	9663	0	0	0
RO	TGJIU	45.03	23.27	203	32819	8038	13529	0	0	0
RO	BUZAU	45.13	26.85	97	27643	6775	11412	0	0	0
RO	DROBE	44.63	22.63	77	33012	8069	13559	0	0	0
RO	BUCUR	44.52	26.08	90	23664	5819	9754	0	0	0
RO	CALAR	44.2	27.33	19	33306	8159	13741	0	0	0
RU	SORTA	61.72	30.72	19	19928	4911	8232	19928	4911	8232
RU	LENIN	59.97	30.3	6	35898	8770	14776	36009	8877	14885
RU	PSKOV	57.82	28.42	45	22076	5457	9147	22077	5458	9148
RU	KALIN	54.72	20.55	21	19175	4735	7922	19198	4732	7934
RU	SMOLE	54.75	32.07	239	20074	4946	8281	20286	5005	8390
RU	KURSK	51.77	36.17	247	35037	8438	14281	35341	8605	14492
SK	HURBA	47.88	18.2	115	18980	4680	7852	18980	4680	7852
ES	SSEBA	43.31	-2.04	259	25896	6387	10716	0	0	0
ES	NAVAC	40.78	-4.01	1890	16058	3960	6643	0	0	0
ES	SALAM	40.95	-5.49	790	17882	4410	7397	0	0	0
ES	BADAJ	38.88	-6.83	185	16425	4050	6795	0	0	0
ES	VALEN	39.48	-0.38	11	23020	5611	9444	0	0	0
ES	TORTO	40.82	0.49	48	22953	5698	9571	0	0	0
SE	VAXJO	56.87	14.8	166	29835	7349	12321	0	0	0
SE	LINKO	58.4	15.53	93	25043	6179	10358	0	0	0
SE	KARLS	59.35	13.47	46	29838	7348	12320	0	0	0
SE	OSTER	63.18	14.48	376	29808	7318	12290	0	0	0
SE	STENS	65.07	17.15	325	29828	7349	12321	0	0	0
CH	BASEL	47.55	7.58	316	36135	8910	14949	36135	8910	14949
CH	SANTI	47.25	9.35	2490	36135	8910	14949	36135	8910	14949
CH	ZURIC	47.38	8.57	556	36135	8910	14949	36135	8910	14949
CH	LUGAN	46	8.97	273	36135	8910	14949	36135	8910	14949
YU	BEOGR	44.8	20.47	132	15330	3780	6342	15317	3769	6329
YU	NIS__	43.33	21.9	202	15330	3780	6342	15330	3780	6342

Table 2.1: Continued

an artifical number calculated from a couple of observations. Unfortunately, a weak spot is present—inhomogenity in methodology. It is universally accepted that mean temperture is a weighted average of several measurements at given times but the times themselves and the weighting varies across Europe and the world. Therefore, any conclusions driven from mean temperatures have to be considered with this inhomogenity in mind. Even without this issue, the principal distinction between mean and minimum/maximum temperatures is still noteworthy.

2.2 Data Preparation

As the raw values may not suitable for testing, data had to be examined first. There are two basic issues to be taken care of:

- Temperatures are subject to trend and seasonal influences. Temperatures of adjacent days are also correlated.
- Data can be polluted by a human error, usually a typo.

Firstly, all measurements from February 29 were discarded to avoid problems with different length of the years and significantly smaller sample size for this particular date.

The data was then checked by simple conditions of min $\leq$ mean $\leq$ max. It was however noted, that the methodology does not guarantee that abovementioned inequalities are fulfilled[2], so suspect dates were reviewed manually. Close mismatches were considered regular, the larger differences were examined to see whether the value is reasonable or a mistake. The other clue was a comparison of first-order differences, suspect values over 20 degrees Celsius were also reviewed and judged in the context of neighbouring values. Several

[2]Min/Max temperature is in fact a temperature measured at a set time, rather than min/max reached in a day-long interval. Likewise, the mean temperature is a weighted mean of a relatively small number of measurements made throughout the day - usually three but the methods vary as noted in the previous chapter.

suspect values were discarded and considered a missing value. Dataset from Vilnius, Lithuania was extraordinarily rich in such suspect values.

As for trend, it is universally accepted in climatology, that a linear trend is assumed in temperatures. Time series (of absolute temperatures) for each day of the year was fitted into a linear trend model. Surprisingly, the data does not support such a premise. Trend coefficient was not significantly different from zero at most stations — with datasets from all stations pooled together, around 90 % were not significant at the 95% significance level. The trend coefficient was also frequently negative (nearly 40 % of datasets) as could be seen from the table on the enclosed CD. Therefore, no adjustment for trend was made.

To remove seasonal influences, the data from each station were grouped by date. Every measurement was then converted from absolute value into relative value compared to the mean temperature on the given day. As the variance of the temperatures also differs by season, differences are given in standard deviations, rather than degrees Celsius.

$$x_i^{norm} = \frac{x_i - \overline{x}}{s}$$

where $\overline{x}$ and s are the sample mean and sample variance on the given day.

Plotting the transformed data shows that there is no obvious periodical pattern (unlike the obvious sinuosity in non-trasformed values).

For better understanding of the difference in temperature behaviour between seasons, data are also examined in smaller seasonal batches. Summer (June, July, August) and winter (December, January, February) are examined separately. Two artificial seasons called “longsummer” (May through September) and “longwinter” (November through March) serve both to overcome possible sample size issues, as well as a security check that the results do not go crazy following small change in the input.

Correlation is another issue. There is a strong correlation between the temperatures of the consecutive days. Previous studies show that higher-order autocorrelation is weak and an AR(1) model is satisfactory. This suggests studying the first-order differences as they can be considered independent, which is an important premise of the normality tests. To verify this, AR model was fitted to both original data and first-order diferences. This time, the original assumption holds as the first-order coefficient of the AR model is at around 0.8 for original data (for mean temperatures, it is actually very close to 1). For diferences, this is cut down to around 0.2, although the higher-order partial autocorrelations go up, which is undesirable.

It needs to be noted that for transformed data, the first-order AR coefficient goes down to around 0.6 (due to the removal of the seasonal effect). Taking first-order differences of the transformed data yields similar results as differences of the non-transformed data. There is no obvious improvement in using diferences of the transformed data versus the differences of the non-transformed data so there is no reason to use this more artifical setting. Differences are therefore the most reliable entity to examine. However, some attention still needs to be paid to the transformed data as these are easier to interpret than the differences.

The histograms (using Scott method of determining the number of bins as Sturges formula produces too wide bins) and kernel denstity estimates (using the default “gaussian” method in R) of the transformed data show that the temperature distribution is often either skewed or bimodal (sometimes even multimodal). On the other hand, most stations visually conform to normal distribution as long as the differences are examined—very few stations exercise an obvious departure from normality. Maximum temperature differences in Verona, Italy produced an extraordinary result that can be disregarded as an obvious singularity.

Chapter 3

Normality Tests

3.1 Description of the Tests Used

Seven basic normality tests were selected from the tests described in [11]. D'Agostino tests are based on the third and fourth moments. There are three different tests that are sensitive to different departures from normality - one sensitive to skewness, one sensitive to kurtosis and the omnibus test that combines both criteria. Four goodness-of-fit tests were also used. Lilliefors test is based on empirical distribution function—it is an extension of Kolmogorov–Smirnov test for the composite hypothesis of normality. Pearson's chi-squared test is used mainly for completeness, although it is less powerful than other tests when testing normality and it is generally not recommended for use to test this specific situation. The last two tests, Anderson–Darling test and Cramer–von Mises test, are also based on empirical distribution function but use a more complicated test statistic than Kolmogorov–Smirnov test. Comparative power study of goodness-of-fit tests was performed in [10].

3.1.1 D'Agostino Skewness Test

The skewness test is based on test statistic

$$a_3 = \frac{m_3}{m_2^{3/2}}$$

where $m_k = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^k}{n}$ is k-th sample moment.

D'Agostino transformation Z_3 is distributed N(0,1) under the null hypothesis. Let

$$U_3 = a_3/\text{var}(a_3) = a_3 \left(\frac{(n+1)(n+3)}{6(n-2)} \right)^{1/2}$$
$$B_2 = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$
$$W^2 = \sqrt{2(B_2-1)} - 1$$
$$\delta = 1/\sqrt{\log(W)}$$
$$\alpha = \sqrt{2/(W^2-1)}$$

Then

$$Z_3 = \delta \log \left(\frac{U_3}{\alpha} + \sqrt{\left(\frac{U_3}{\alpha} \right)^2 + 1} \right)$$

For the values that $|Z_3| > u(\frac{\alpha}{2})$, the hypothesis of normality is rejected by the D'Agostino skewness test.

3.1.2 D'Agostino Kurtosis Test

The kurtosis test is based on test statistic

$$a_4 = \frac{m_4}{m_2^2}$$

Let

$$U_4 = \frac{a_4 - Ea_4}{\text{var}(a_4)}$$
$$B = \sqrt{\frac{216}{n}} \left\{ \frac{(n+3)(n+5)}{(n-3)(n-2)} \right\}^{1/2} \frac{n^2-5n+2}{(n+7)(n+9)}$$

$$A = 6 + \frac{8}{B}\left(\frac{2}{B} + \sqrt{1 + 4/B^2}\right)$$

Then

$$Z_4 = (2/9A)^{-\frac{1}{2}}\left\{1 - \frac{2}{9A} - \left(\frac{1-2/A}{1+x\sqrt{2/(A-4)}}\right)^{1/3}\right\}$$

is once again N(0,1) distributed under the null hypothesis. D'Agostino kurtosis test rejects normality for the values so that $|Z_4| > u(\frac{\alpha}{2})$.

3.1.3 D'Agostino Omnibus Test

Omnibus test combines the two above-mentioned tests in one. Under the null hypothesis, the test statistic

$Z = Z_3^2 + Z_4^2$

is χ_2^2 distributed. Normality is rejected for the values so that $Z > \chi_2^2(\alpha)$.

3.1.4 Lilliefors Test

The Kolmogorov-Smirnov test statistic is the maximum distance between empirical distributional function (EDF) and distribution function under the null hypothesis. Let

$\mathrm{p}_{(i)} = \Phi([x_{(i)} - \overline{x}]/s)$

Then, Lilliefors test statistic is defined as

$\mathrm{D} = \max(D^+, D^-)$

$\mathrm{D}^+ = \max_{i=1,\ldots,n}[i/n - p_{(i)}]$,

$\mathrm{D}^- = \max_{i=1,\ldots,n}[p_{(i)} - (i-1)/n]$

The Lilliefors test rejects the hypothesis of normality for $D > D_\alpha$.

3.1.5 Anderson–Darling Test

Anderson–Darling test is also based on EDF, the test statistic being

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1][\log(p_{(i)}) + \log(1 - p_{(n-i+1)})]$$

The p-value is computed from the Stephens' modified statistic

$A_*^2 = (1 + 0.75/n + 2.25/n^2)A^2$

The Anderson–Darling test rejects the hypothesis of normality for $A_* > A_*(\alpha)$.

3.1.6 Cramer–von Mises Test

Cramer–von Mises test is a close relative to the Anderson–Darling test, with a test statistic

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(p_{(i)} - \frac{2i-1}{2n} \right)$$

The p-value is again computed from the Stephens' modified statistic

$W_*^2 = (1 + 0.5/n)W^2$

The Cramer–von Mises test rejects the hypothesis of normality for $W_* > W_*(\alpha)$.

3.2 Single Normality Tests

The following maps show p-values of a test performed on data from a single given station. Generally, it is obvious that it is insufficient to model temperature by normal distribution as at most stations the hypothesis of normality is rejected even at the 99% confidence level. However, several noteworthy patterns are recognized. Because of obvious space limitation, not all maps mentioned could be shown here. All maps are available on the enclosed CD.

Normality tests presume an independent data. Two sets of data are examined. "Standardized data" refers to temperatures transformed by the method explained in the previous chapter. Due to rather strong correlation in the time series, these results are very unprecise and can only give a very rough guide. "Differences" refer to first-order differences of the non-transformed data and are the main focus of this chapter as the correlation does not have such a strong effect.

The resulting p-values are colour-coded—black dots denote stations that reject the test at the 99% confidence level. Coloured stations do not reject the test at the 99% level—p-values under 0.05 are red (the test is rejected at the 95% level), p-values under 0.1 are green, (dark) blue denotes a p-value of 0.1–0.5 and finally cyan[1] denotes a p-value greater than 0.5.

Despite the fact that the maps that are shown in the text are black-and-white (with a full-color version avaiable on the CD), any station that do not reject normality at the 99% level can be referred to as a "coloured" station as opposed to a black station.

There was an issue with Cramer–von Mises test that often returned p-values greater than one. Simulations showed that for sufficiently large samples ($n > 1800$) coming from uniform distribution, large values of the test statistic (corresponding to p-values very close to zero) were converted to p-values much greater than one. This seems to be a technical error in the way R software computes p-values, rather than malfunction of the test itself. Therefore all p-values greater than one are treated as small numbers very close to zero.

3.2.1 Mean Temperature

Mean temperatures generally show lesser fit to normality than minimum and maximum temperatures. D'Agostino omnibus test is the prime example as normality is not rejected at the 99% level only at few stations across all seasons. However, both skewness and kurtosis tests produce more stations that pass the test, compared to omnibus, indicating different types of departures. For skewness, there are only a few stations that do not reject the test for full year data (most are located on the North/Batlic Sea coast, see Figure 3.1). Seasonal data show some logical patterns—in summer the non-rejecting stations are in the Scandinavia and Iceland, for winter, such stations are found in central Europe. In the light of this, the full year results might indicate a boundary between the two regions with different behaviour where the two influences even out (see Figure 3.2). Sample skewness values were examined with regard to the sign and the result supports this reasoning. The sample skewness is generally negative in summer and positive in winter, ie. it is more common to see a large drop in temperature than a large rise in summer

[1]Light blue

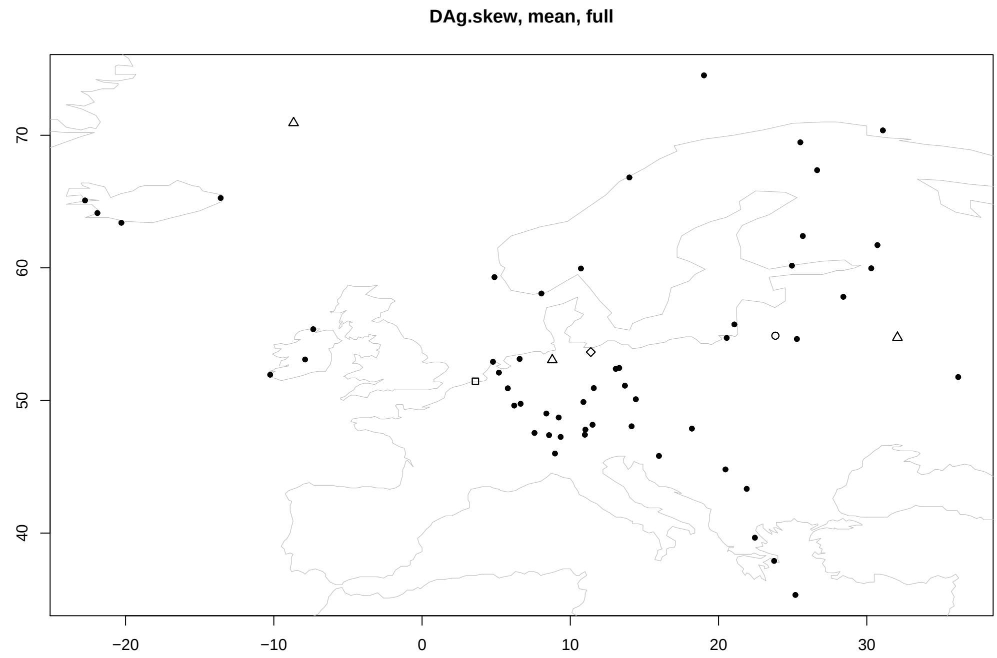

Figure 3.1: D'Agostino skewness, mean temperature, full year: Stations that pass the test are located around the border visible in Figure 3.2. Legend: p-value is (dot) below 1%, (circle) below 5%, (square) below 10%, (diamond) below 50%, (triangle) above 50%

and the other way around in winter. At the first sight, this might sound natural, as warming would be expected in winter, but the seasons have the warmest/coldest point in the middle. To verify this, seasonal densities were examined as well as seasonal sample means and no systematic flaw was found. In fact, the sample means were frequently positive in summer and negative in winter, which is a direct opposite of the abovementioned reasoning.

The other trend in sample skewness values is that positive values are found in the north and west, while negative values are concentrated in the south and east. Full year values are pretty balanced in terms of number of positive/negative stations but the regional grouping is obvious and the stations passing the test (close enough to zero) are located at the border. This region happens to be the one where the winter and summer behaviour

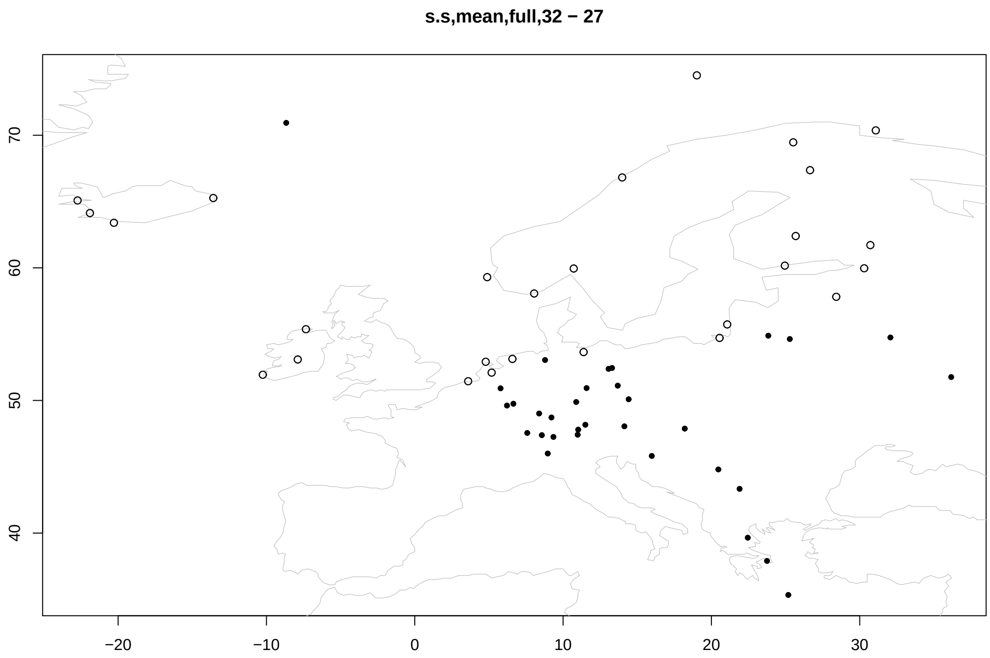

Figure 3.2: Sample skewness, mean temperature, full year: Positive and negative sample skewness is clearly tied to latitude. Legend: Full dot: negative; empty circle: positive; Headline numbers: negative–positive count

evens out.

Situation is much clearer with kurtosis, as almost all stations produced higher kurtosis (lighter tails) than a normal distribution,[2] regardless of the season.

Other tests do not show such interesting results, there are few stations that do not reject hypothesis of normality.

Looking at the standardized data, the sample skewness behaves similarly as for differences, but this time summer produces positive values, winter negative values, and full year is a mixture of both with positive values located northwards, although the two regions are not separated as sharply as for differences. Kurtosis seems to be influenced by sea distance rather than latitude, with all the positve stations located at the coast for full year. The

[2]For convenience, negative/positive will refer to lower/greater than three.

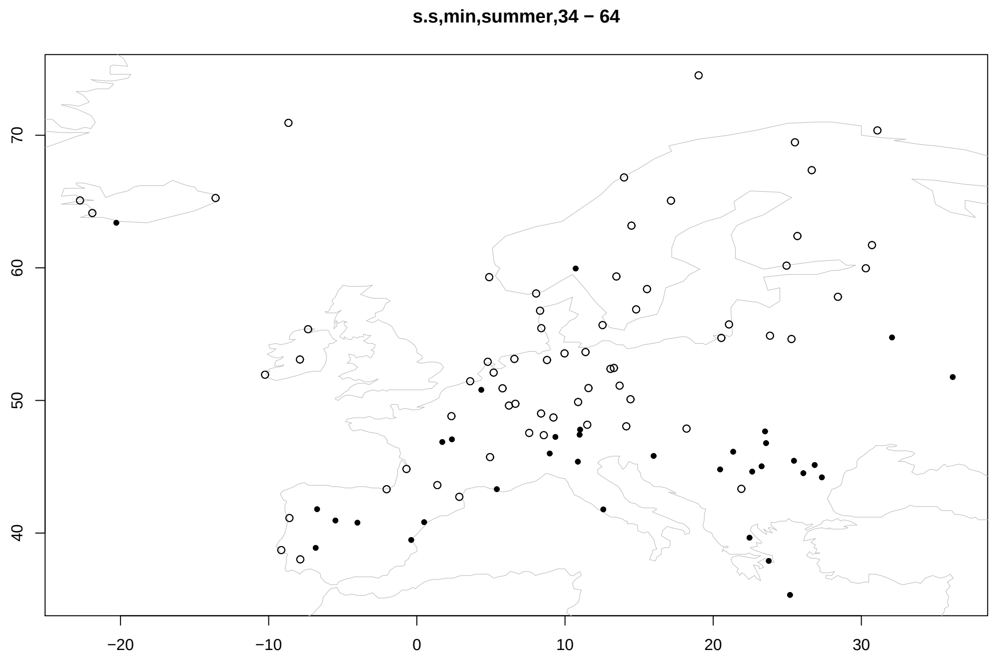

Figure 3.3: Sample skewness, minimum temperature, summer: Stations are split southeast vs. northwest. Legend: Full dot: negative; empty circle: positive; Headline numbers: negative–positive count

size of the two groups is more balanced but the sign changes between winter and summer at majority of the stations—summer is similar to the full year but in winter, the positive stations are in the middle of the map and the negative stations are all around them.

3.2.2 Minimum Temperature

The most notable feature of skewness is a prevalence of positively-skewed stations in winter (this corresponds well with the mean temperatures). There are scattered stations that do not reject normality but no pattern seems to be present apart from the fact that they tend to be in the south. For summer, positive stations are still more common (unlike with mean temperatures) with the divison being southeast (negative) vs. northwest (positive). A tie between the boundary and the non-rejecting stations is not particularly strong in this case. Obviously, most stations have positive sample skewness for full year results. Romania is the only region with common negative stations, the same region that had at least some negative stations in winter. The few stations that happen to pass the normality test seem to be random occurences.

Kurtosis is again larger than 3, ocassionaly even as large as 11-12 (summer in Romania), although there are several stations where normality is not rejected in summer (and to lesser extent in longsummer).

Omnibus test has only a few scattered stations that pass the test in summer. This is in accordance with the large winter kurtosis and the presence of several non-rejections in summer for both moments. The goodness-of-fit tests give no useful results. Hardly any station passes the test for any season.

Similar to mean temperatures, sample skewness and kurtosis of the standardized data is usually of the opposite sign than for differences. Negative skewness is prevalent in winter (this also transfers to full year), summer is more balanced with positive skewness in the northwestern half of Europe (although Iceland and Ireland are negative, see Figure 3.3). Summer kurtosis is (likewise mean temperatures) positive in coastal stations. In winter, positive kurtosis is found in the central part of Europe, with negative results around it—once again the same pattern as for mean temperatures.

3.2.3 Maximum Temperature

Nothing new with kurtosis—sample values are unanimously positive for maximum temperatures in all seasons. Winter skewness is divided regionally, with positive values in the north and negative in the south—a trend that corresponds with summer minimum temperatures, although southwest is now the base of the negative region (while it was southeast for summer minimum temperatures). In summer, most stations produce negative skewness but the northern stations are positive and the normality test is not rejected at a couple of Scandinavian stations, suggesting a continuous rise in the sample skewness on the way north (see Figure 3.4). Likewise, many stations that do

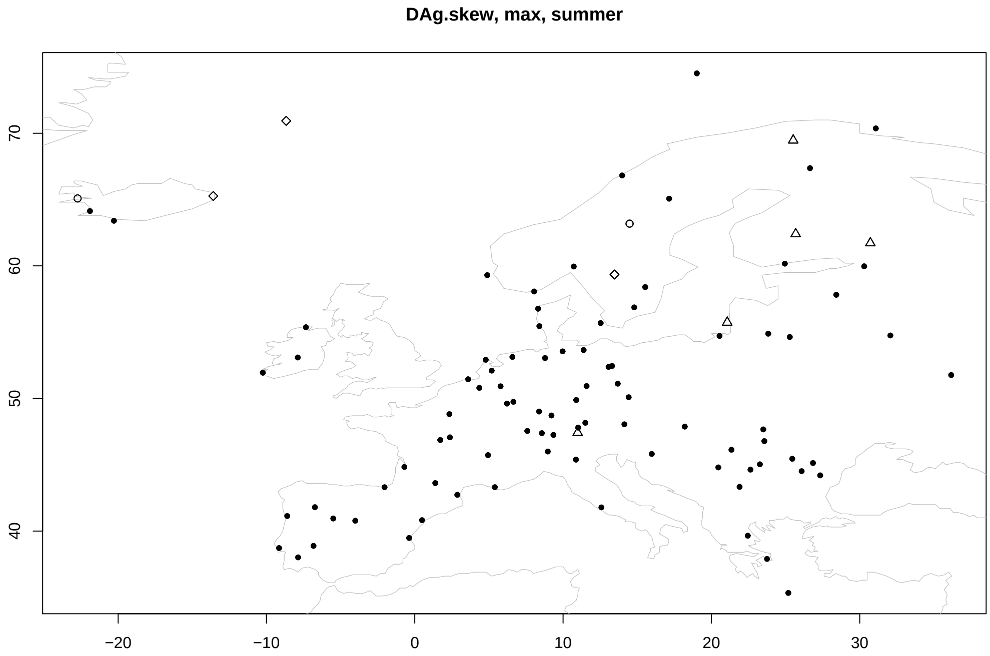

Figure 3.4: D'Agostino skewness, maximum temperature: In summer, several Scandinavian stations do not reject normality. Legend: p-value is (dot) below 1%, (circle) below 5%, (square) below 10%, (diamond) below 50%, (triangle) above 50%

not reject normality in winter are located in the area where the positive and negative regions meet. The goodness-of-fit tests are unanimously rejected.

Looking at the standardized data, sample skewness sign stays negative and the northwest–southeast division still applies like it did for minimum temperatures. In summer, sample kurtosis remains different in coastal and landlocked areas (see Figure 3.5). In winter, numbers are pretty balanced with positive and negative stations mixed together. Also, D'Agostino kurtosis test is quite often not rejected, so the values are often close to the normal distribution kurtosis.

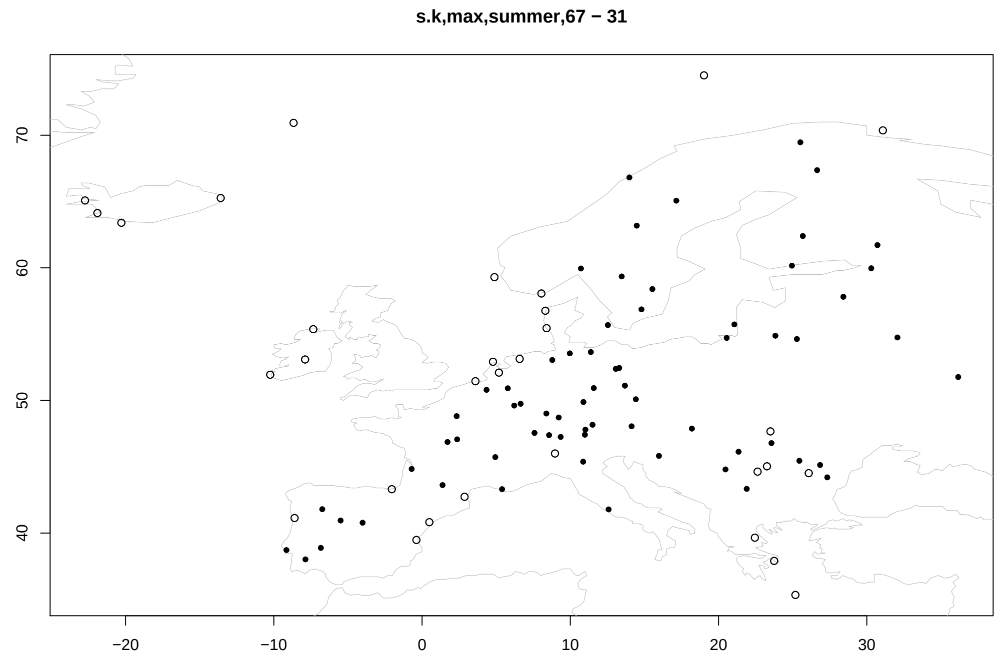

Figure 3.5: Standardized data, sample kurtosis, maximum temperature, summer: The sign seems to be related to the sea distance. Legend: Full dot: negative; empty circle: positive; Headline numbers: negative–positive count

3.3 Multiple Test Comparison

It is quite important for interpretation of the results whether the same stations keep passing all the different tests. Every single test result could be just a random fluke and a few such occasions in one region could appear as a pattern to human observer. In the following figures, number of tests that did not reject normality at the 99% level are shown. These results are disscussed unless mentioned otherwise. Results obtained at the 95% level were also examined, but the non-rejecting stations are already quite rare for the 99% level.

With mean temperatures, only winter and longwinter has two stations that do not reject multiple tests—Malin, Ireland and Sodankylä, Finland. Sodankyla is also the only such station in summer seasons but the for the full year data, all tests are rejected.

For minimum temperatures, summer is the season with the most non-rejections and there are two regions—France and Germany—with several such stations in one place. However, German stations often pass two tests only (D'Agostino kurtosis and Engelmann-Hartigan). Jyväskylä, Finland and Smolensk, Russia are the two stations that have no such neighbours. Other seasons are not interesting, even the longsummer season does not replicate the summer results and only Germany is somewhat notable.

For maximum temperatures, only Zugspitze, Germany passed all three D'Agostino tests in summer, other results are not interesting.

Chapter 4

Finite Normal Mixtures

As the normal distribution is not sufficient to model temperature, a family of finite normal mixture distributions, that is discussed in Thode [11], will be examined. A k-component univariate normal density is defined as

$$h(x) = \sum_{i=1}^{k} \pi_i \phi(x; \mu_i, \sigma_i^2)$$

where $\phi(x)$ denotes a normal density and $\sum_{i=1}^{k} \pi_i = 1$.

This distribution family has a nice feature—it includes a wide range of shapes, asymmetric as well as symmetric. Although one could expect the mixtures to be multimodal, unimodal distrubutions are obviously present as well. Skewed distrubutions are also covered. All in all, finite normal mixtures seem to be a good candidate for improving the temperature modelling.

Finite normal mixtures will be examined in two ways. The Engelman–Hartigan test [2] is a normality test against an alternative of a two-component normal mixture with equal variances. The other approach utilises the EM algorithm to find the best finite normal mixture model available.[1]

[1] Normal distribution could be viewed as single-component normal mixture and therefore normal distributions are also considered.

4.1 Engelman–Hartigan Test

4.1.1 Description of the Test

The Engelman–Hartigan test is a test of normality against a two-component normal mixture with equal variances. The observations $x_1, x_2, \ldots, x_n$ are divided into two groups of size n_1 and n_2 ($n_1 + n_2 = n$) and the ratio of between to within sum of squares is calculated as

$$SS_B/SS_W = \frac{n_1 n_2 (\overline{x}_1 - \overline{x}_2)^2}{(n_1 + n_2)[\sum (x_i - \overline{x}_1)^2 + \sum (x_i - \overline{x}_2)^2]}$$

where $\overline{x}_j$ is the mean of group j and the sums are over the respective groups.

Engelman–Hartigan test statistic is

$$EH = \max(SS_B/SS_W)$$

The maximum is computed over all possible divisions. The maximum can only occur for one of the $n-1$ divisions of ordered data, so only these actually need to be calculated. Although an argument could be made that low values of EH (that suggest lack of randomness in the data) speak against the hypothesis too, usually the one-sided version of the test is used. The hypothesis of normality is therefore rejected for $EH > c_\alpha$, where c_α is the critical value at the α level of significance. For c_α, the following approximation was found empirically.

$$\log(c_\alpha + 1) = -\log(1 - 2/\pi) + (n-2)^{-1/2} z_{1-\alpha} + 2.4(n-2)^{-1}$$

where $z_{1-\alpha}$ is the percentile of the standard normal distribution. In other words, the following transformation of EH

$$U = \sqrt{n-2}\,[\log(EH + 1) + \log(1 - 2/\pi) - 2.4/(n-2)]$$

is N(0,1) distributed.

It is important to note that the Engelman–Hartigan test has a specific alternative so the results are not reliable if the true distribution is different from the postulated distribution. Non-rejections tend to show a presence

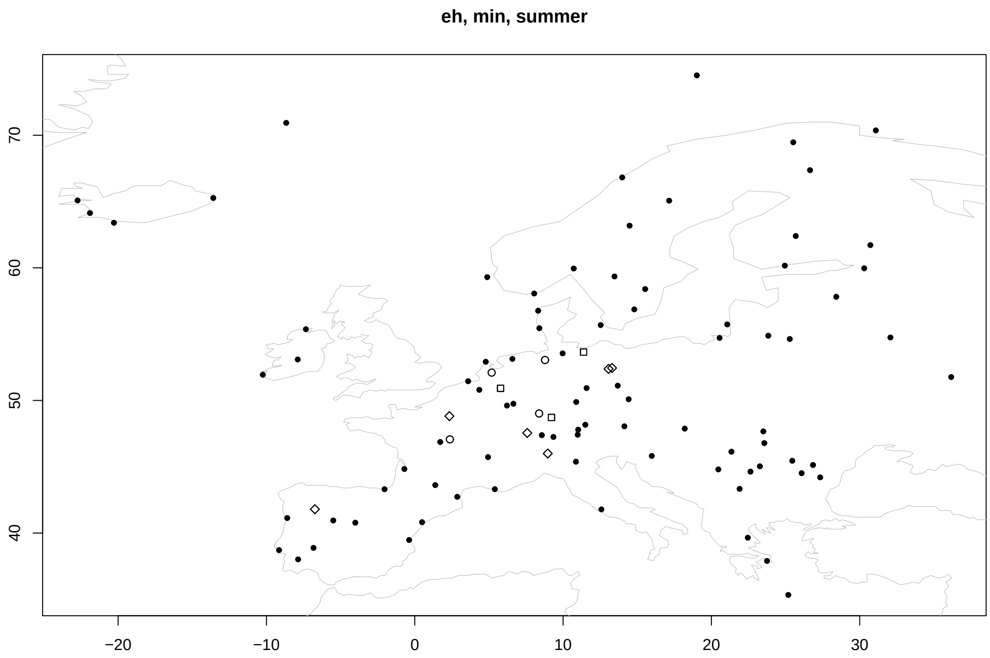

Figure 4.1: Engelman–Hartigan, minimum temperatures, summer: Germany and the surroundings form a region where the test is not rejected. Legend: p-value is (dot) below 1%, (circle) below 5%, (square) below 10%, (diamond) below 50%, (triangle) above 50%

of a single phenomenon rather than a solid proof of normality.[2] On the other hand, the results of this test should match the results given by the EM algorithm.

4.1.2 Test Results

Firstly, the Engelman–Hartigan test results are plotted in the same style as the normality tests in the previous chapter. Starting with the mean temperatures, hardly any stations do not reject the test, a few exceptions are usually in the summer. This is also the only season where a region with frequent non-rejections is found in the south of central Europe (longsummer

[2]D'Agostino skewness and kurtosis test behave in a similar way.

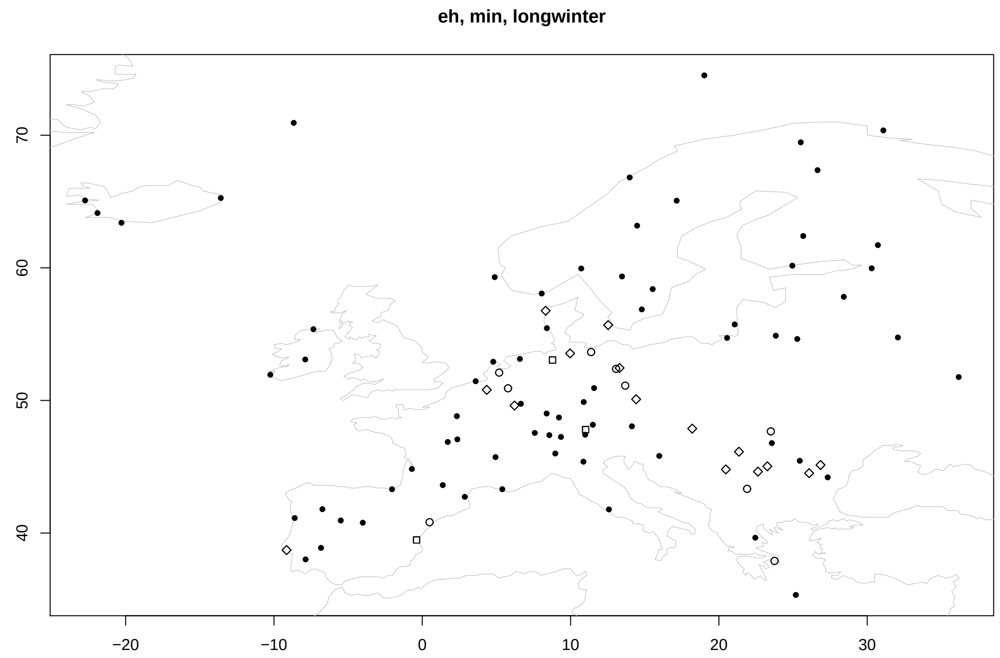

Figure 4.2: Standardized data: Engelman–Hartigan, minimum, longwinter: Non-rejecting stations are located between the Benelux and the Balkans.
Legend: p-value is (dot) below 1%, (circle) below 5%, (square) below 10%, (diamond) below 50%, (triangle) above 50%

season copies the region to some extent but the area is visibly smaller). For minimum temperatures, such a region is found in Germany and its western and southern neighbourhood (see Figure 4.1). For maximum temperatures, this region moves to the east to central Europe and the Balkans and retains about the same size. In the winter seasons and the full year, there are just rejections with a single exception.

Standardized data give more colourful results. For mean temperature, winter has a number of coloured stations in the central Europe and the Balkans. For summer, such stations are less frequent and located at the sea, usually in the north. On the other hand, the full year results are rather non-descript and no pattern is present. For minimum temperatures, many northern stations do not reject the test in summer but almost all do in winter. In the rest of Europe, more stations are rejected in summer but no real patterns stand out. Curiously enough, the longwinter map resembles the mean temperature situation (see Figure 4.2),but the winter season itself does not. Maximum temperatures seem to give the least useful result—in summer, a few scattered stations do not reject the test; in winter, such stations tend to be in the southeast but many rejecting stations are spread among them.

4.2 EM Algorithm

4.2.1 Description of the Method

EM algorithm[3] is a two-step iterative procedure that finds the best model for a given number of components and variance model [3]. For the univariate data, there are just two models—"E" for equal variance throughout all components and "V" for each component having a different variance.[4] Bayesian Information Criterion (BIC) is computed as an indicator of quality of the model. The BIC value corresponds to the model type and number of components as well, so it is comparable across all models. The model with maximum BIC value is then selected as the best finite normal mixture model.

A disadvantage of this method is a great computational intensity. The iterative process is quite slow and larger datasets are difficult to handle. To overcome this limitation, the data for each station were split into several parts, the EM algorithm was applied at each part separately, and the results were compared. In this study, three groups are formed. The maximum possible subsample size is 4200 observations. If the sample size allows to create three distinct subsets of size 4200, the first 4200 and last 4200 observations form two groups, with the third one comprised of the 4200 observations around the middle mark. Otherwise, the sample is split in thirds of equal size. As the sample size of short seasons is lower, another set of computations was made for groups of size 2000 so that the comparisons across the seasons

[3]EM stands for "Expectation and Maximization" steps of the algorithm

[4]In the single-component case, this distinction does not exist and the model is labeled as "X".

could be made with similar sample sizes too. This number was chosen on the basis of short seasons' sample size but it is also considered adequate size for the EM algorithm, as further increasing of the sample size does not improve the results much [4].

Both standardized data and differences were examined with this method—the former are more practical, the comparison of the results with the previous chapter will be interesting for the latter.

During the first computations, it become apparent that it is not necessary to compute models of five or more components, as 4-component models were selected as best very exceptionally, and higher numbers never showed up. Therefore, maximum number of components for the algorithm was set to four, saving a lot of resources for what should be a negligible price.

4.2.2 Result Presentation

Three subsets of data were examined for each station, selecting the best model for each subset. The three models are compared and the result is presented in the map. For the purpose of this presentation, the following convention is used:

- Large gray dot denotes a station where three different models were selected.
- If at least two results match, an empty circle shows that the best model has equal variances, while small full dot denotes a model with different variances. For single component, an empty square is plotted.
- Colour displays the number of components. Black/Cyan stands for single component, Red/Magenta for two, Green/Yellow for three and Blue/Gray for four components.
- First colour (RGB + Black) denotes that all three subsets agreed, while the other (CMY + Gray) indicates that one subset arrived at different

model than the other two subsets.[5]

As an auxiliary criterion, the same map is also produced exclusively on the basis of number of components to help in deciding how serious a partial mismatch is. If the odd subset produces model with the same number of components as the other two, credibility of the model is somewhat increased.

4.2.3 Standardized Data

For the mean temperatures, a two-component mixture with equal variances is the most common result, particularly in summer. Unequal variance models can be found in colder climate (Scandinavia, Northern/Baltic Sea coast). A 3-component mixture with equal variances is present in a small area in Germany. The most common alternative is the normal model that is common in the south (and somewhat suprisingly in Iceland) in the summer. For full year data, this model is widespread in central Europe but some normal modelled stations are found everywhere across the map. The situation is different in winter, as this is the season where three-component mixtures are rather common—mostly in the cold regions (Finland, Russia). Unlike summer, 3-component mixtures rarely have equal variances in winter. Otherwise, all different models can be found—normal models are found in the south but the two-component mixtures with equal variances are mixed among them. This model is also the most common in the central Europe and as the climate gets colder, the model first changes to unequal variances and then to the three-component mixture.

Minimum temperatures for the full year data are again a combination of the normal and 2-component (equal variances) model. The normal model tends to be more frequent in the south, but the two models are mixed up next to each other. There is also a rather large disagreement between the results from the different sample sizes as the larger samples produce a rather compact region of the 2-component mixture with unequal variance in north-east (Scandinavia and Russia), while this model is only occasionally selected

[5]The best model is judged only on the basis of number of components and the variance type. The value of parameters is not considered here.

with the smaller samples. Still, this result supports the general trend of the gradual increase of the model complexity on the way from mild to harsh climate.

Summer minimum temperatures are a mix of the normal and 2-component model with equal variance and this time the normal model appears routinely in the north too as the two models are intervowen. Normal models seem to be more frequent in the coastal areas but it is not a strict rule. Winter is again the most colourful season with all models present—normal in the Mediterranean, 2-component-equal around it, 2-component-unequal in the central Europe and 3-component in Scandinavia. Some unusual patterns are found too, such as the 3-component stations in France.

Maximum temperature full year results are basically split between normal and 2-component-equal-variance models with little regard to latitude as the normal model is present at all regions. For summer, the usual northeast-to-southwest division is modified to the northwest-southeast direction. Two-component models with unequal variance are dominant in the northeast but some three-component stations are present there too. Winter is the usual mix of 3-component models in Scandinavia with 2-component-unequal models surrounding (although these are uncharacteristically found in France) and the rest is split between 2-component-equal and normal models.

4.2.4 Differences

For mean temperatures, a 2-component mixture is by far the most common model for all seasons. Stations with different results are only singularities—normal models are found in the north, as well as 3-component mixtures (one such result is also found in winter at an Alpine station but the smaller sample result disagreed). Of the 2-component models, equal variances are quite rare, with one exception being the landlocked stations in central Europe in summer, where this model clearly prevails.

The two-component mixture with unequal variances is the prevalent model in minimum temperatures too. Usual alternatives are found in winter—three-component mixtures in the north and normal models in the south.

In summer, the normal model is the most common and the two-component mixtures are spread all across Europe. For full year, the two sample sizes produce different results. Small samples show a couple of regions with a different model than the widespread 2-component-unequal—normal model in France and surroundings, equal variances in central Europe and a couple of 3-component stations. For the larger samples, 3-component stations are more common, but the other alternatives are almost non-existant.

Maximum temperatures produced the most stable results. Vast majority of stations follows a 2-component model with unequal variances in all seasons. Scandinavia is the only notable exception. Three-component model is present in winter (and also for the full year). There are two surprises there, however. The longwinter season has more 3-component stations than winter. In summer (and longsummer to lesser extent) the normal model is present, rather than the 3-component.

4.3 Comparison of Results

In order to compare the results of the Engelman–Hartigan test and EM algorithm, the results of the two approaches are compared, observing whether the test returns the same result on a given station. All results are then summarized in a 2×2 contingency table. Fisher test for independence is then applied to see if there is a connection. This test is described e.g. in Anděl [1]. The one-sided alternative of $\delta > 0$ is used as we a priori expect the results to match. In the maps presented, “F.test” states the p-value of the (one-sided) Fisher test and “OR” is the point estimate of the odds ratio. Before drawing any conclusions from the p-values, an important feature of Fisher test needs to be noted. The discrete nature of the situation implies that the possible resulting p-values are also discrete and especially with low-count cells of the table, the p-values are unstable in a sense that a small change of data earns a big leap in the p-value.

The results of the Engelman–Hartigan test are divided into two groups based on the p-values—one group rejects the test at the given level, the

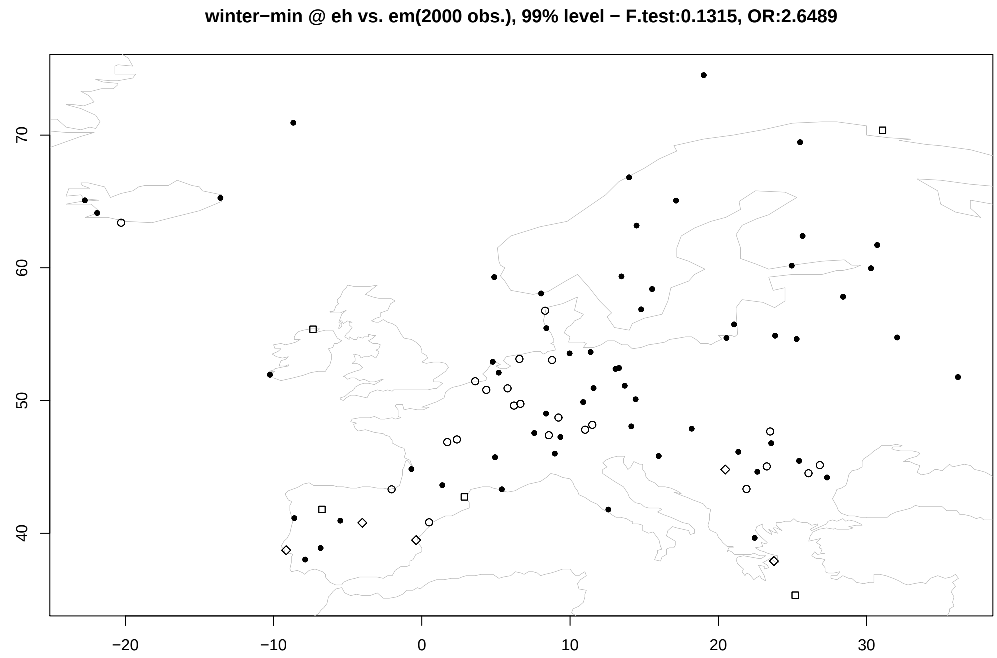

Figure 4.3: Standardized data, minimum temperature (one-sided): Winter produces the worst fit of the two tests which might be blamed upon the higher ratio of the 3-component stations. Legend: (dot) Both tests "rejected" normality, (circle) E.–H. test rejected, (square) EM algorithm "rejects" normality, (diamond) both tests imply normality.

other does not. The decision point was once again set at the 95% and 99% level.[6] As for the EM algorithm, stations that return normal model are treated as non-rejections, stations with a 2-component model are treated as rejections. Gray dots (complete mismatches) and 3-component stations need to be accomodated as well. The most obvious solution is to leave out mismatches and treat 3-component stations as rejections, i.e. the same way as 2-component stations. Other possibilities[7] were briefly examined as well. Leaving out the 3-component stations tends to increase the p-values in winter

[6]The 95% level is more common but because of small observed p-values, the 99% level is useful for the decent-sized samples in this study.

[7]Treating both issues the same way—either rejecting or leaving out.

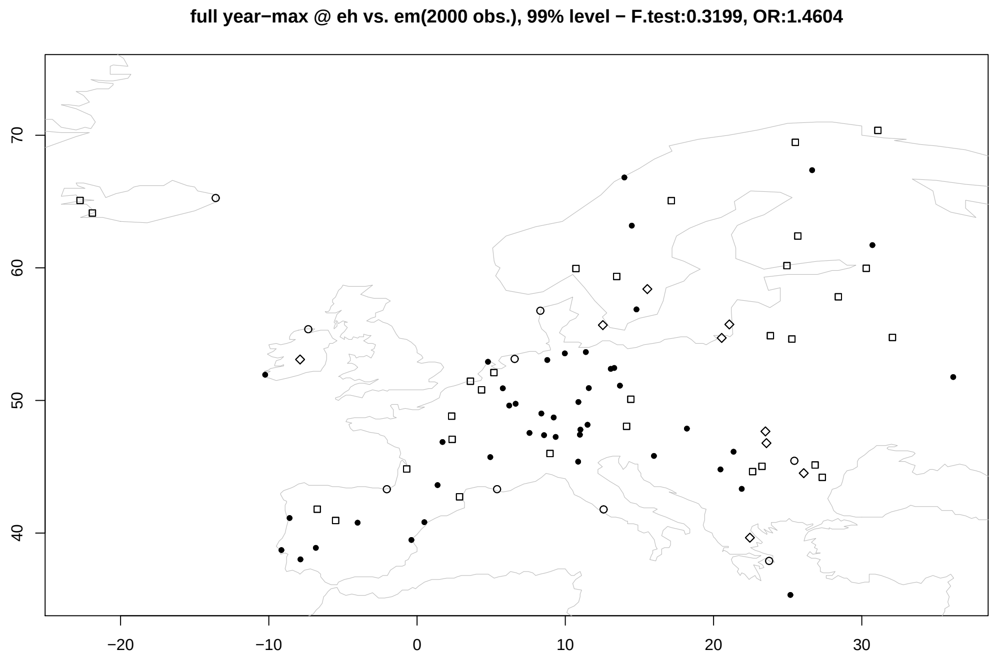

Figure 4.4: Standardized data, maximum temperature, full year (one-sided): Smaller sample often selected normality in Scandinavia. Legend: (dot) Both tests "rejected" normality, (circle) E.–H. test rejected, (square) EM algorithm "rejects" normality, (diamond) both tests imply normality.

(a seson with the highest frequency of such stations). However, the difference is not dramatic, so only the default method is discussed.

For the standardized data, the results of the comparison are quite satisfactory, especially using the 95% level of the Engelmann–Hartigan test. The odds ratio estimate was always larger than one. All p-values are summed up in Table 4.1. Minimum/maximum temperature results match very well as the independence of the two tests is often rejected even at the 99% significance level. Winter (and longwinter) minimum temperatures are not rejecting the independence test (see Figure 4.3) but this is not a major problem, as this is the situation with the highest frequency of a three-component model, so the Engelmann–Hartigan test could be expected to be the least accurate of all conditions examined. The full year maximum temperature p-value is a big-

min	mean	max	99% vs. 95%	min	mean	max
0.0015	0.1323	0.3199	**full**	0.0013	0.1166	0.5368
0.0003	0.0856	0.0039	**summer**	0.0026	0.0406	0.0654
0.1315	0.1626	0.0859	**winter**	0.0576	0.2420	0.3778
0.0050	0.0894	0.0044	**longsummer**	0.1424	0.3300	0.0684
0.2037	0.0882	0.0004	**longwinter**	0.3728	0.6646	0.0069

Table 4.1: Fisher test p-values for standardized data. The left-hand table corresponds to the Engelman–Hartigan test at the 99% level. The right-hand part is for the 95% level.

ger surprise (see Figure 4.4). It is a consequence of the high frequency of the normal models in the EM-algorithm. Mean temperature results might somewhat suffer from a smaller number of stations and therefore a smaller sample size for the Fisher test because of the discrete nature of the test. It is however safe to claim that the mean temperatures do not fit the assumption of the finite normal mixture distribution as well as the minimum/maximum temperatures. Using the EM-algorithm results obtained with the larger samples does not dramatically change the results—at least for the seasonal data. The p-values tend to be somewhat lower, on the other hand the low values often go up a little. Full year p-values go significantly down for mean temperatures (below 5 % in both cases) and for the full year results (to approximately half of the tabulated value). The latter is the effect of the difference spotted in EM-algorithm results, but such an obvious explanation is not avaiable for the former.

Diferences are not very useful to examine because the Engelman-Hartigan test rejected most of the stations. For summer and longsummer, the number of non-rejections was acceptable but no strong connection between the two methods could be estabilished.

On the whole, the results of this comparison support the finite normal mixture model for the temperature data, so the results of the EM-algorithm can be used with confidence.

Chapter 5

Conclusion

It is clearly demonstrated that the presumption of normality in temperatures does not hold universally and serious attention has to be paid to the circumstances if a statistical method is based on such a presumption. The temperature distribution depends heavily on the time of the year and location, as well as the temperature type. The mean temperature behaves differently than the minimum and maximum temeprature. The latter two also exhibit major differences between warm and cold periods of the year.

Daily first-order temperature differences were examined and although the goodness-of-fit tests do not show normality, some useful results were obtained at least for the higher order moments. Using finite normal mixtures offers a possible improvement over normality. Cross-checking comparing normal model and the 2-component normal mixture showed a significant correspondence between the normality test against this specific alternative and the results of the EM-algorithm that validates using finite normal mixtures as an alternative model. A two-component normal mixture is usually enough to improve the model, unless the climate conditions are unusual. A three-component normal mixture handles the severe skewness of winter temperatures in the north Europe as it manages to accomodate increased number of very cold days. The normal models are often adequate in the south Europe—in the sense that finite normal mixtures do not improve the model as the distribution is close to normal, although the EDF goodness-of-fit tests do not recommend treating the distribution as such for the purpose of methods sensitive to the normality of data.

Apart from a full collection of maps, the enclosed CD also contains full results of the tests, i.e. p-values of normality tests and the details of the models selected by the EM algorithm (especially all the parameters), that could be helpful for an analysis tied to a certain location.

Bibliography

[1] Jiří Anděl. *Základy matematické statistiky.* UK MFF, Prague, 2002. Preprint.

[2] L. Engelman and J.A. Hartigan. Percentage points of a test for clusters. *Journal of the American Statistical Association*, 64:1647–1648, 1969.

[3] C. Fraley and A. E. Raftery. MCLUST:software for model-based clustering, density estimation and discriminant analysis. Technical report, Department of Statistics, University of Washington, 2002. See URL: http://www.stat.washington.edu/mclust.

[4] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[5] Dao-Yi Gong and Chang-Hoi Ho. Intra-seasonal variability of wintertime temperature over East Asia. *International Journal of Climatology*, 24:131–144, 2004.

[6] R.D. Harmel, C.W. Richardson, C.L. Hanson, and G.L. Johnson. Evaluating the adequacy of simulating maximum and minimum daily air temperature with the normal distribution. *Journal of Applied Meteorology*, 41:744–753, 2002.

[7] Radan Huth, Jan Kyselý, and Martin Dubrovský. Simulation of surface air temperature by GCMs, statistical downscaling and weather generator: Higher–order statistical moments. *Stud. Geophys. Geod.*, 47:203–216, 2003.

[8] A.M.G. Klein Tank and Coauthors. Daily dataset of 20th-century surface air temperature and precipitation series for European Climate Assessment (ECA), 2001. Data and metadata available at http://www.knmi.nl/samenw/eca or http://www.knmi.nl/samenw/ecd.

[9] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[10] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.

[11] Henry C. Thode Jr. *Testing for normality*. Dekker, New York, USA, 2002. ISBN 0-8247-9613-6.

[12] Zoltan Toth and Tamas Szentimrey. The binormal distribution: A distribution for representing asymmetrical but normal–like weather elements. *Journal of Climate*, 3:128–136, 1990.