Influence of network density on homogenization performance

José A. Guijarro (jguijarrop@aemet.es)

State Meteorological Agency, Palma de Mallorca, Spain

VII Seminar for Homogenization and Quality Control in Climatological Databases (Budapest, 24-28 October 2011)
Outline

- Motivation
- Methodology:
  - Network generation
  - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Outline

- Motivation
- Methodology:
  - Network generation
  - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Outline

- Motivation
- Methodology:
  - Network generation
    - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Outline

- Motivation
- Methodology:
  - Network generation
  - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Outline

- Motivation
- Methodology:
  - Network generation
  - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Outline

- Motivation
- Methodology:
  - Network generation
  - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Outline

- Motivation
- Methodology:
  - Network generation
  - Tested homogenization programs
- Results and discussion
- Conclusions
- Future work
Motivation (a)

Homogenization comparison has been problematic in COST Action ES0601:

- **Complex structure of the benchmark data-set tree:**
  - Manual methods could homogenize only a reduced subset
  - Automatic methods had often post-processing errors
- **High number of random inhomogeneities and local trends**
- **Method performance difficult to compare**
Motivation (a)

Homogenization comparison has been problematic in COST Action ES0601:

- Complex structure of the benchmark data-set tree:
  - Manual methods could homogenize only a reduced subset
  - Automatic methods had often post-processing errors
- High number of random inhomogeneities and local trends
- Method performance difficult to compare
Homogenization comparison has been problematic in COST Action ES0601:

- Complex structure of the benchmark data-set tree:
  - Manual methods could homogenize only a reduced subset
  - Automatic methods had often post-processing errors
- High number of random inhomogeneities and local trends
- $\Rightarrow$ Method performance difficult to compare
Motivation (a)

Homogenization comparison has been problematic in COST Action ES0601:

- Complex structure of the benchmark data-set tree:
  - Manual methods could homogenize only a reduced subset
  - Automatic methods had often post-processing errors
- High number of random inhomogeneities and local trends
  - ⇒ Method performance difficult to compare
Motivation (a)

Homogenization comparison has been problematic in COST Action ES0601:

- Complex structure of the benchmark data-set tree:
  - Manual methods could homogenize only a reduced subset
  - Automatic methods had often post-processing errors
- High number of random inhomogeneities and local trends
- \(\Rightarrow\) Method performance difficult to compare
Motivation (b)

- Benchmark data-set tried to be highly realistic
- Yet it was composed by long series only, with few missing data
- ⇒ Climatol’s ability to use nearby short series remained untested
- **Objective:** To evaluate the impact of using these short series
Motivation (b)

- Benchmark data-set tried to be highly realistic
- Yet it was composed by long series only, with few missing data
- Climatol’s ability to use nearby short series remained untested
- **Objective:** To evaluate the impact of using these short series
Motivation (b)

- Benchmark data-set tried to be highly realistic
- Yet it was composed by long series only, with few missing data
- → Climatol’s ability to use nearby short series remained untested

Objective: To evaluate the impact of using these short series
Motivation (b)

- Benchmark data-set tried to be highly realistic
- Yet it was composed by long series only, with few missing data
- \( \Rightarrow \) Climatol’s ability to use nearby short series remained untested
- **Objective:** To evaluate the impact of using these short series
Methodology (a)

Base network generation:

- 100 random points on a 4 x 3° lon-lat area
- Seasonal cycle taken from monthly averages of maximum daily temperatures of 53 stations from the central area of the Duero river basin, Spain ($t_i$, °C):
  
  7.9 10.5 14.2 16.3 20.6 25.8 29.9 29.3 25.2 18.7 12.2 8.3

- First station: 60 years of random monthly values taken from $N(t_i, 1.5)$

- Closest station to any with already assigned values: Add to the closest values $p \cdot N(0, 1.5)$, with $p = 0.20, 0.40, 0.80$ (3 basic networks were generated: TA20, TA40, TA80)

- Continue until all 100 stations have been filled with values

- All series are then adjusted for altitude, their annual range is varied by -20 to 20%, and a trend of 2°C/Century is added.
Methodology (a)

Base network generation:

- 100 random points on a 4 x 3° lon-lat area
- Seasonal cycle taken from monthly averages of maximum daily temperatures of 53 stations from the central area of the Duero river basin, Spain ($t_i$, °C):
  
  | 7.9 | 10.5 | 14.2 | 16.3 | 20.6 | 25.8 | 29.9 | 29.3 | 25.2 | 18.7 | 12.2 | 8.3 |

- First station: 60 years of random monthly values taken from $\mathbf{N}(t_i, 1.5)$
- Closest station to any with already assigned values: Add to the closest values $p \cdot \mathbf{N}(0, 1.5)$, with $p = 0.20, 0.40, 0.80$ (3 basic networks were generated: TA20, TA40, TA80)
- Continue until all 100 stations have been filled with values
- All series are then adjusted for altitude, their annual range is varied by -20 to 20%, and a trend of 2°C/Century is added.
Methodology (a)

Base network generation:

- 100 random points on a 4 x 3° lon-lat area
- Seasonal cycle taken from monthly averages of maximum daily temperatures of 53 stations from the central area of the Duero river basin, Spain ($t_i$, °C):
  
  7.9 10.5 14.2 16.3 20.6 25.8 29.9 29.3 25.2 18.7 12.2 8.3
- First station: 60 years of random monthly values taken from $\mathbf{N}(t_i, 1.5)$
- Closest station to any with already assigned values: Add to the closest values $p \cdot \mathbf{N}(0, 1.5)$, with $p = 0.20, 0.40, 0.80$
  (3 basic networks were generated: TA20, TA40, TA80)
- Continue until all 100 stations have been filled with values
- All series are then adjusted for altitude, their annual range is varied by -20 to 20%, and a trend of 2°C/Century is added.
Methodology (a)

Base network generation:

- 100 random points on a 4 x 3° lon-lat area
- Seasonal cycle taken from monthly averages of maximum daily temperatures of 53 stations from the central area of the Duero river basin, Spain ($t_i, ^\circ C$):
  
  7.9 10.5 14.2 16.3 20.6 25.8 29.9 29.3 25.2 18.7 12.2 8.3

- First station: 60 years of random monthly values taken from $\mathbf{N}(t_i, 1.5)$
- Closest station to any with already assigned values: Add to the closest values $p \cdot \mathbf{N}(0, 1.5)$, with $p = 0.20, 0.40, 0.80$ (3 basic networks were generated: TA20, TA40, TA80)
- Continue until all 100 stations have been filled with values
- All series are then adjusted for altitude, their annual range is varied by -20 to 20%, and a trend of 2°C/Century is added.
Methodology (a)

Base network generation:

- 100 random points on a 4 x 3° lon-lat area
- Seasonal cycle taken from monthly averages of maximum daily temperatures of 53 stations from the central area of the Duero river basin, Spain ($t_i$, °C):
  
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.9</td>
<td>10.5</td>
<td>14.2</td>
<td>16.3</td>
<td>20.6</td>
<td>25.8</td>
<td>29.9</td>
</tr>
<tr>
<td>29.3</td>
<td>25.2</td>
<td>18.7</td>
<td>12.2</td>
<td>8.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- First station: 60 years of random monthly values taken from $N(t_i, 1.5)$
- Closest station to any with already assigned values: Add to the closest values $p \cdot N(0, 1.5)$, with $p = 0.20, 0.40, 0.80$ (3 basic networks were generated: TA20, TA40, TA80)
- Continue until all 100 stations have been filled with values

- All series are then adjusted for altitude, their annual range is varied by -20 to 20%, and a trend of 2°C/Century is added.
Methodology (a)

Base network generation:

- 100 random points on a 4 x 3° lon-lat area
- Seasonal cycle taken from monthly averages of maximum daily temperatures of 53 stations from the central area of the Duero river basin, Spain ($t_i$, °C):
  
<table>
<thead>
<tr>
<th>7.9</th>
<th>10.5</th>
<th>14.2</th>
<th>16.3</th>
<th>20.6</th>
<th>25.8</th>
<th>29.9</th>
<th>29.3</th>
<th>25.2</th>
<th>18.7</th>
<th>12.2</th>
<th>8.3</th>
</tr>
</thead>
</table>

- First station: 60 years of random monthly values taken from $N(t_i, 1.5)$
- Closest station to any with already assigned values: Add to the closest values $p \cdot N(0, 1.5)$, with $p = 0.20, 0.40, 0.80$ (3 basic networks were generated: TA20, TA40, TA80)
- Continue until all 100 stations have been filled with values
- All series are then adjusted for altitude, their annual range is varied by -20 to 20%, and a trend of 2°C/Century is added.
Correlogram of first difference series

Correlogram of TA20 basic network
Correlogram of first difference series

Correlogram of TA40 basic network
Correlogram of first difference series

Correlogram of TA80 basic network
Methodology (b)

- 40 stations randomly chosen from every basic network in 100 different runs
- 2°C shifts in the mean imposed on the first 5 stations
- Following 5 stations unchanged (homogeneous references)
- Stations 11 to 40 keep only between 10 and 30 years of data (17 to 50%)
Methodology (b)

- 40 stations randomly chosen from every basic network in 100 different runs
- 2°C shifts in the mean imposed on the first 5 stations
- Following 5 stations unchanged (homogeneous references)
- Stations 11 to 40 keep only between 10 and 30 years of data (17 to 50%)
Methodology (b)

- 40 stations randomly chosen from every basic network in 100 different runs
- $2^\circ$C shifts in the mean imposed on the first 5 stations
- Following 5 stations unchanged (homogeneous references)
- Stations 11 to 40 keep only between 10 and 30 years of data (17 to 50%)
Methodology (b)

- 40 stations randomly chosen from every basic network in 100 different runs
- 2°C shifts in the mean imposed on the first 5 stations
- Following 5 stations unchanged (homogeneous references)
- Stations 11 to 40 keep only between 10 and 30 years of data (17 to 50%)
Example test series
Methodology (c)

Tested homogenization programs:
- **ACMANT** (Domonkos) + wine: ’Acm’
- **Climatol** (Guijarro): 10, 20 and 40 stations, with constant (’cl1’, ’cl2’, ’cl4’) and variable (’Cl1’, ’Cl2’, ’Cl4’) corrections
- **RHTTestV2** (Wang): Using as reference the mean of the other 9 stations (’RH1’) or of the 5 homogeneous stations only (’RH2’)
- **HOMER** (Mestre et al.) + expect: r.min=0.5, min.neighbors=5, d (pairwise d.) + j (joint d.) + c (correction); ’Hom’
- **Pending:** MASH (Szentimrey) and AnClim/ProClimDB (Stépánek), both through wine
Methodology (c)

Tested homogenization programs:

- ACMANT (Domonkos) + wine: 'Acm'
- Climatol (Guijarro): 10, 20 and 40 stations, with constant ('cl1', 'cl2', 'cl4') and variable ('Cl1', 'Cl2', 'Cl4') corrections
- RHTestV2 (Wang): Using as reference the mean of the other 9 stations ('RH1') or of the 5 homogeneous stations only ('RH2')
- HOMER (Mestre et al.) + expect: r.min=0.5, min.neighbors=5, d (pairwise d.) + j (joint d.) + c (correction); 'Hom'
- Pending: MASH (Szentimrey) and AnClim/ProClimDB (Stépánek), both through wine
Tested homogenization programs:

- **ACMANT (Domonkos) + wine:** ‘Acm’
- **Climatol (Guijarro):** 10, 20 and 40 stations, with constant (‘cl1’, ‘cl2’, ‘cl4’) and variable (‘Cl1’, ‘Cl2’, ‘Cl4’) corrections
- **RHTestV2 (Wang):** Using as reference the mean of the other 9 stations (‘RH1’) or of the 5 homogeneous stations only (‘RH2’)
- **HOMER (Mestre *et al.*):** expect: r.min=0.5, min.neighbors=5, d (pairwise d.) + j (joint d.) + c (correction); ‘Hom’
- **Pending:** MASH (Szentimrey) and AnClim/ProClimDB (Stépánek), both through wine
Tested homogenization programs:

- ACMANT (Domonkos) + wine: ‘Acm’
- Climatol (Guijarro): 10, 20 and 40 stations, with constant (‘cl1’, ‘cl2’, ‘cl4’) and variable (‘Cl1’, ‘Cl2’, ‘Cl4’) corrections
- RHTestV2 (Wang): Using as reference the mean of the other 9 stations (‘RH1’) or of the 5 homogeneous stations only (‘RH2’)
- HOMER (Mestre et al.) + expect: r.min=0.5, min.neighbors=5, d (pairwise d.) + j (joint d.) + c (correction); ’Hom’

Pending: MASH (Szentimrey) and AnClim/ProClimDB (Stépánek), both through wine
**Methodology (c)**

Tested homogenization programs:

- **ACMANT** (Domonkos) + wine: ’Acm’
- Climatol (Guijarro): 10, 20 and 40 stations, with constant (’cl1’, ’cl2’, ’cl4’) and variable (’Cl1’, ’Cl2’, ’Cl4’) corrections
- RHTestV2 (Wang): Using as reference the mean of the other 9 stations (’RH1’) or of the 5 homogeneous stations only (’RH2’)
- **HOMER** (Mestre *et al.*) + expect: r.min=0.5, min.neighbors=5, d (pairwise d.) + j (joint d.) + c (correction); ’Hom’
- **Pending**: MASH (Szentimrey) and AnClim/ProClimDB (Stépánek), both through wine
Results and discussion

- The series are homogenized taking the last homogeneous period as reference
- Four statistical parameters have been evaluated for comparing the five homogenized series with the originals:
  1. RMSE (main measure of homogenization performance)
  2. Difference of trends (climate change detection)
  3. Difference of means (climate mapping)
  4. Difference of standard deviations (climate variability...)
Results and discussion

- The series are homogenized taking the last homogeneous period as reference.
- Four statistical parameters have been evaluated for comparing the five homogenized series with the originals:
  1. RMSE (main measure of homogenization performance)
  2. Difference of trends (climate change detection)
  3. Difference of means (climate mapping)
  4. Difference of standard deviations (climate variability...)
Results and discussion

- The series are homogenized taking the last homogeneous period as reference
- Four statistical parameters have been evaluated for comparing the five homogenized series with the originals:
  1. RMSE (main measure of homogenization performance)
  2. Difference of trends (climate change detection)
  3. Difference of means (climate mapping)
  4. Difference of standard deviations (climate variability...)
Results and discussion

- The series are homogenized taking the last homogeneous period as reference
- Four statistical parameters have been evaluated for comparing the five homogenized series with the originals:
  1. RMSE (main measure of homogenization performance)
  2. Difference of trends (climate change detection)
  3. Difference of means (climate mapping)
  4. Difference of standard deviations (climate variability...
Results and discussion

- The series are homogenized taking the last homogeneous period as reference.

- Four statistical parameters have been evaluated for comparing the five homogenized series with the originals:
  1. RMSE (main measure of homogenization performance)
  2. Difference of trends (climate change detection)
  3. Difference of means (climate mapping)
  4. Difference of standard deviations (climate variability...).
Results and discussion

- The series are homogenized taking the last homogeneous period as reference.
- Four statistical parameters have been evaluated for comparing the five homogenized series with the originals:
  1. RMSE (main measure of homogenization performance)
  2. Difference of trends (climate change detection)
  3. Difference of means (climate mapping)
  4. Difference of standard deviations (climate variability...)
TA80 Std. dev. differences (°C)

Methods

Inh  Acm  cl1  cl2  cl4  Cl1  Cl2  Cl4  RH1  RH2  Hom

−1.0  −0.5  0.0  0.5  1.0

TA80 Std. dev. differences (°C)
TA20 Mean differences (°C)

Methods

Mean differences (°C)
TA40 Mean differences (°C)

Methods

Inh, Acm, cl1, cl2, cl4, Cl1, Cl2, Cl4, RH1, RH2, Hom
TA40 Trend differences (°C/100 years)
TA20, TA40 and TA80 Trend differences
TA20 RMSE (°C)
Conclusions

- Impact of short series only noticeable with low correlations
- Climatol performance good even without the help of short series
- The major benefit of the ability to process the short series is for climate applications (mapping, monitoring, ...)
- Automatic benchmarking helps testing new advances in homogenization methods
Conclusions

- Impact of short series only noticeable with low correlations
- Climatol performance good even without the help of short series
- ⇒ The major benefit of the ability to process the short series is for climate applications (mapping, monitoring, ...)
- Automatic benchmarking helps testing new advances in homogenization methods
Conclusions

- Impact of short series only noticeable with low correlations
- Climatol performance good even without the help of short series
- The major benefit of the ability to process the short series is for climate applications (mapping, monitoring, ...)
- Automatic benchmarking helps testing new advances in homogenization methods
Conclusions

- Impact of short series only noticeable with low correlations
- Climatol performance good even without the help of short series
- The major benefit of the ability to process the short series is for climate applications (mapping, monitoring, ...)
- Automatic benchmarking helps testing new advances in homogenization methods
Future work

- Include MASH and AnClim/ProClimDB in the tests
- Optimize HOMER’s sequence of computations
- Study the effect of shortening the last homogenous period
- Introduce random and seasonaly dependent shifts in the series (in the first 5, and in all of them)
- Simulate variables with different seasonality (e.g.: precipitation)
- Simulate daily values of temperature and precipitation
- **Final remark:** The scripts used in this work are freely available (just ask for them!)
Future work

- Include MASH and AnClim/ProClimDB in the tests
- Optimize HOMER’s sequence of computations
  - Study de effect of shortening the last homogenous period
  - Introduce random and seasonaly dependent shifts in the series (in the first 5, and in all of them)
  - Simulate variables with different seasonality (e.g.: precipitation)
- Simulate daily values of temperature and precipitation

**Final remark:** The scripts used in this work are freely available (just ask for them!)
Future work

- Include MASH and AnClim/ProClimDB in the tests
- Optimize HOMER’s sequence of computations
- Study the effect of shortening the last homogenous period
- Introduce random and seasonally dependent shifts in the series (in the first 5, and in all of them)
- Simulate variables with different seasonality (e.g.: precipitation)
- Simulate daily values of temperature and precipitation
- **Final remark:** The scripts used in this work are freely available (just ask for them!)
Future work

- Include MASH and AnClim/ProClimDB in the tests
- Optimize HOMER’s sequence of computations
- Study the effect of shortening the last homogenous period
- Introduce random and seasonally dependent shifts in the series (in the first 5, and in all of them)
- Simulate variables with different seasonality (e.g.: precipitation)
- Simulate daily values of temperature and precipitation
- **Final remark:** The scripts used in this work are freely available (just ask for them!)
Future work

- Include MASH and AnClim/ProClimDB in the tests
- Optimize HOMER’s sequence of computations
- Study the effect of shortening the last homogenous period
- Introduce random and seasonally dependent shifts in the series (in the first 5, and in all of them)
- Simulate variables with different seasonality (e.g.: precipitation)
- Simulate daily values of temperature and precipitation
- **Final remark:** The scripts used in this work are freely available (just ask for them!)
Future work

- Include MASH and AnClim/ProClimDB in the tests
- Optimize HOMER’s sequence of computations
- Study the effect of shortening the last homogenous period
- Introduce random and seasonally dependent shifts in the series (in the first 5, and in all of them)
- Simulate variables with different seasonality (e.g.: precipitation)
- Simulate daily values of temperature and precipitation

Final remark: The scripts used in this work are freely available (just ask for them!)
Future work

▶ Include MASH and AnClim/ProClimDB in the tests
▶ Optimize HOMER’s sequence of computations
▶ Study de effect of shortening the last homogenous period
▶ Introduce random and seasonal dependent shifts in the series (in the first 5, and in all of them)
▶ Simulate variables with different seasonality (e.g.: precipitation)
▶ Simulate daily values of temperature and precipitation
▶ Final remark: The scripts used in this work are freely available (just ask for them!)