

The US Temperature Record 5: Preparing the Data

/ SEP 23, 2021

1. Chose the dataset you want to use. I use the raw data because as a chemist I learned to use the data I recorded, and any fiddling you want to do needs to be done in the model. I'll be comparing the data later in this series. As an example, I'll use the tmax.raw data. Grab the file from <https://www.ncei.noaa.gov/pub/data/ushcn/v2.5/> (about 5 MB), which has the .tar.gz extension. Most zip programs can open this, and you'll need to drill into the directory structure (four-levels deep!) and unzip the .txt files in its own directory. I named mine tmax.raw. It contains 1200+ files.
2. Concatenate all the files into a single text file named tmax.raw.txt by first opening a command line window (windows-R, "cmd", enter) and navigate to the folder you just created:

```
cd downloads
cd ushcn
cd tmax.raw
dir (to confirm files are there)
copy *.* Tmax.raw.txt
```

3. This creates a file, Tmax.raw.txt, which is about 18 MB long.
4. Import this data into a database program. I'll use MS Access, but libreBase works also. A spreadsheet won't work too well because we'll need to average all temperature readings for each year. The import function needs to define the columns of the data. Here is the list of the data columns, and there are a lot of them:

Variable -----	Columns -----	Type -----
ID	1-11	Integer
YEAR	13-16	Integer
VALUE1	17-22	Integer
DMFLAG1	23-23	Character
QCFLAG1	24-24	Character
DSFLAG1	25-25	Character
.	.	.
.	.	.
.	.	.
VALUE12	116-121	Integer
DMFLAG12	122-122	Character

QCFLAG12	123-123	Character
DSFLAG12	124-124	Character

Variable Definitions:

ID: 11 character alphanumeric identifier:

- characters 1-2=Country Code ('US' for all USHCN stations)
- character 3=Network code ('H' for Historical Climatology Network)
- digits 4-11='00'+6-digit Cooperative Observer Identification Number

YEAR: 4 digit year of the station record.

VALUE: monthly value (MISSING=-9999). Temperature values are in hundredths of a degree Celsius, but are expressed as whole integers (e.g. divide by 100.0 to get whole degrees Celsius).

- Precipitation values are in tenths of millimeters, but are
- also expressed as whole integers (e.g. divide by 10.0 to
- get millimeters).

5. And repeat this process for any other datasets you want to graph. Create each in a new table of the database. You can save the import format (they call it the "import specification") so the subsequent imports are much easier than setting up the first. Each data table will be about 150,000 rows long, each representing one station's monthly averages for that year. What I did was build one import specification, then used that same "spec." for each dataset. They all have the same format. I'd name the resulting table as whatever it was (T_{\max} -raw, T_{avg} -TOB for example).
6. Now, they didn't make this easy. Whenever a value is not known, they used the value -9999. All those need to be changed to [null] using the search-and-replace function; just leave the new value blank. I needed to do that several times per table to get them all changed.
7. If you want to do what I'm doing by importing every dataset, plan to spend some time at it. The resulting .acddb file will be quite large, 325 MB. Maybe I'll update this file each year and make it available for download. Depends how well it compresses. These don't need to be updated often; the time base for climate change is very long, decades. Annual updates are more than often enough.

Revision #2

Created 7 April 2024 19:09:48 by bruce

Updated 7 April 2024 19:11:18 by bruce