LIVING DATA FOR CLIMATE CHANGE RESEARCH

-GIUDITTA PAROLINI-

SPICED ACADEMY DATA SCIENCE BOOTCAMP

German Phenology data

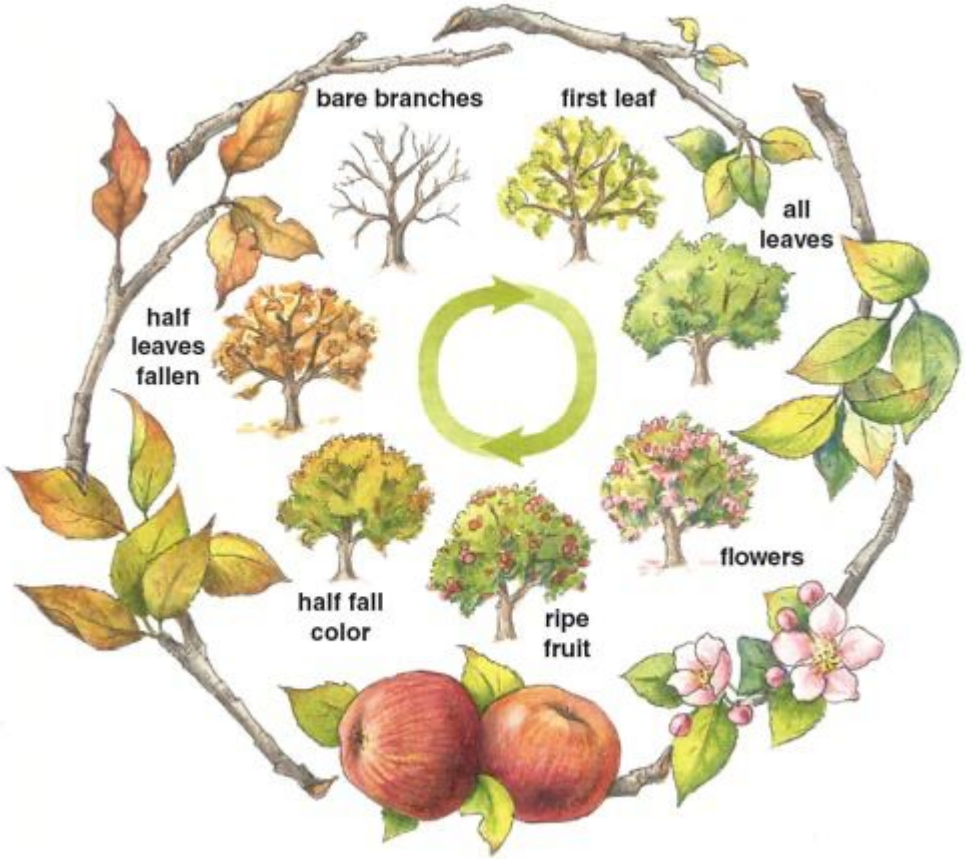
Over 11 million data points for the years 1951-2019.

Data made available by the PEP725 project

Data

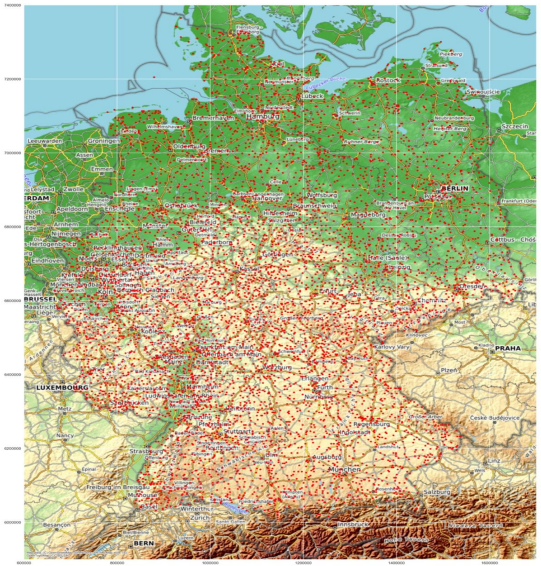
```
s_id                0
provider_id         0
lon                 0
lat                 0
alt                 0
alt_dem             0
gss_id              0
genus               0
species        401226
subspecies   11177192
phase_id            0
year                0
day                 0
date                0
cult_season         0
affected_flag       0
qc_flag             0
qc_ori_flag         0
dtype: int64
```

Missing data



Phenophases for the apple tree (Malus domestica)
https://www.americanscientist.org/article/citizen-science-takes-root

German Phenological stations



TECH STACK

Dask and Pandas (low_memory=False) for building the full dataset, cleaning the data, doing EDA

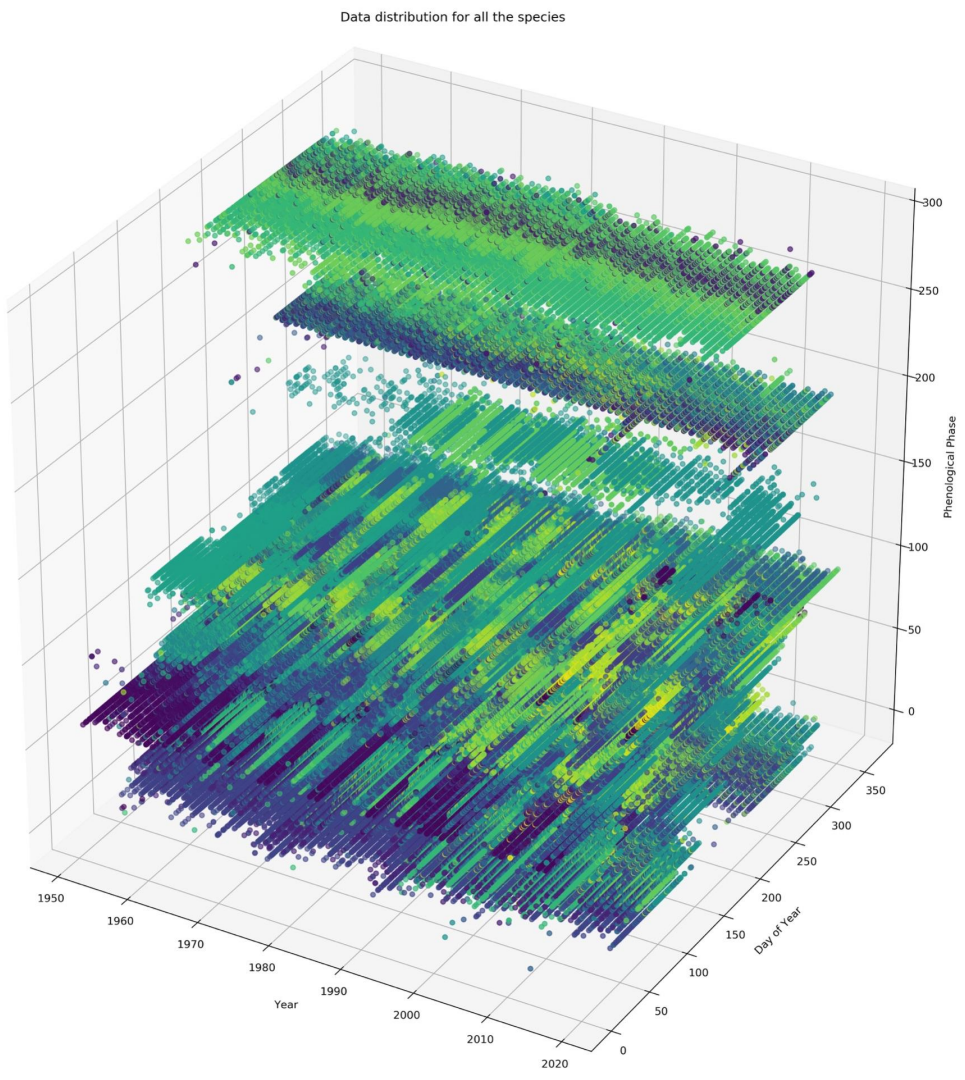
GeoPandas, Folium for making maps

Scikit-Learn and Statsmodels for analysing the data

Phenological species

In total, 55 plant species surveyed in the data set

Variable amount of observations for each species: e.g. Hordeum vulgare (barley) 716.812 observations, Populus tremula (Aspen) 844 observations

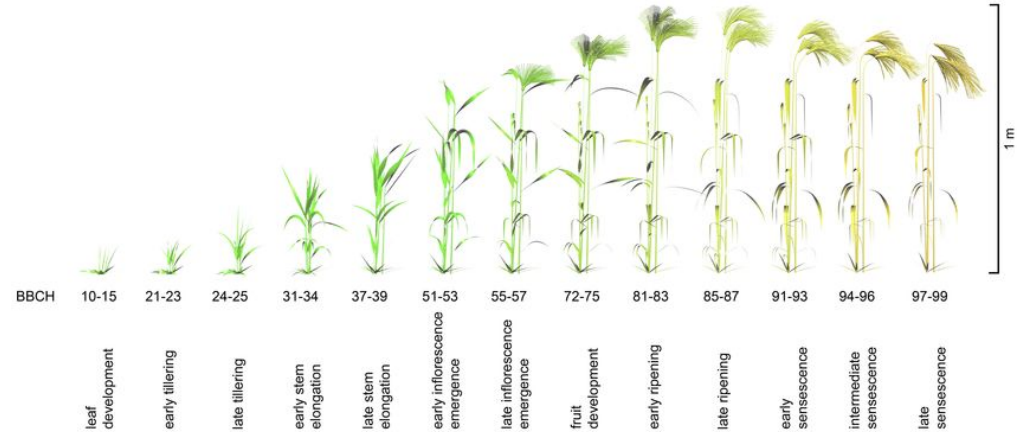


Barley (Hordeum vulgare)

Over 700.000 data points (plant with the highest number of observations in the data set)

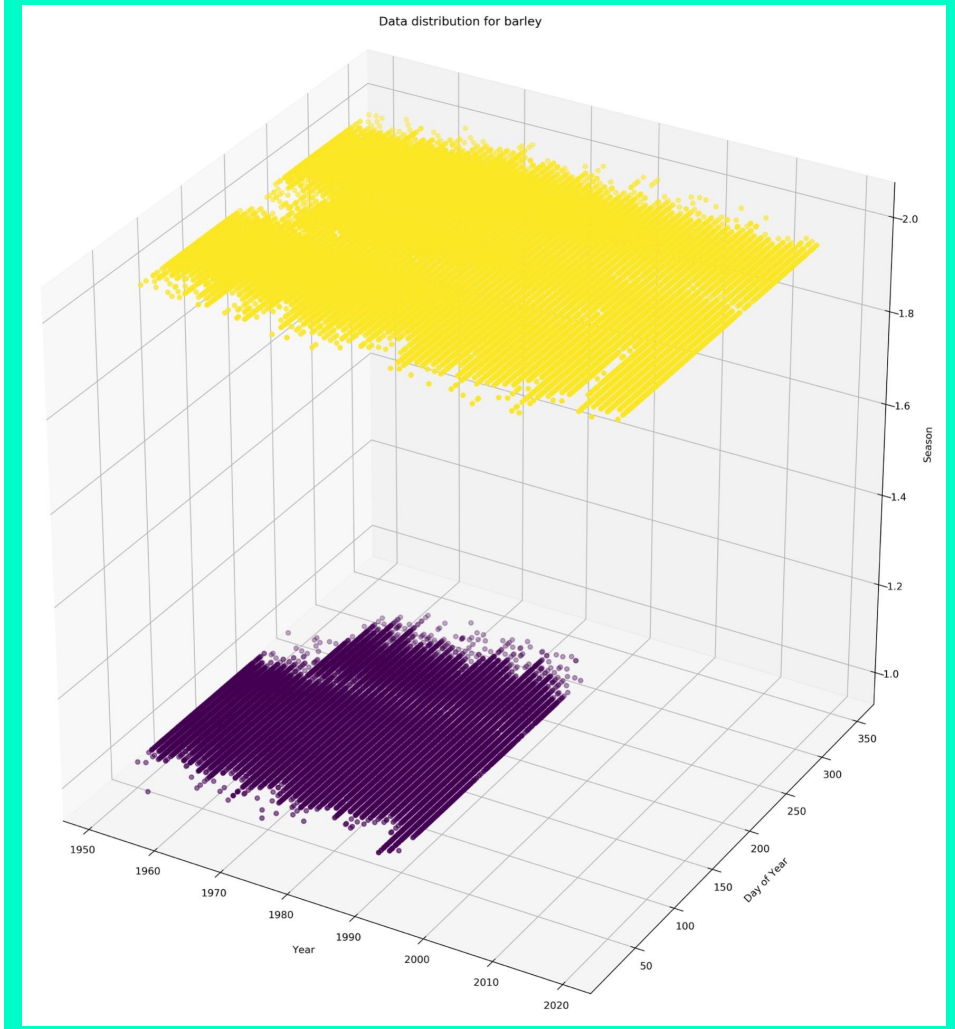
Surveyed plant phases:

- **0**: Dry seed
- **10**: First true leaf emerged from coleoptile
- **31**: One node detectable
- **51**: Beginning of heading
- **85**: Soft dough stage: grain content soft but dry
- **100**: Start of harvest

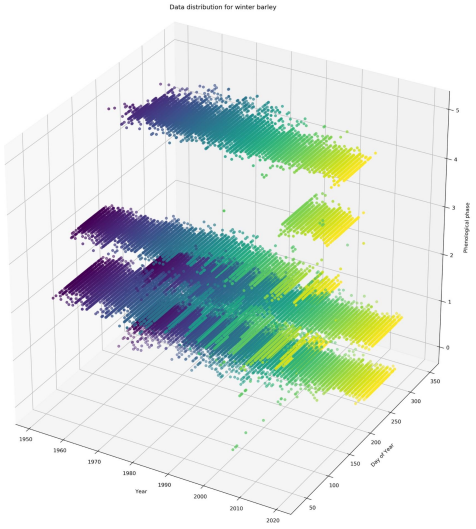


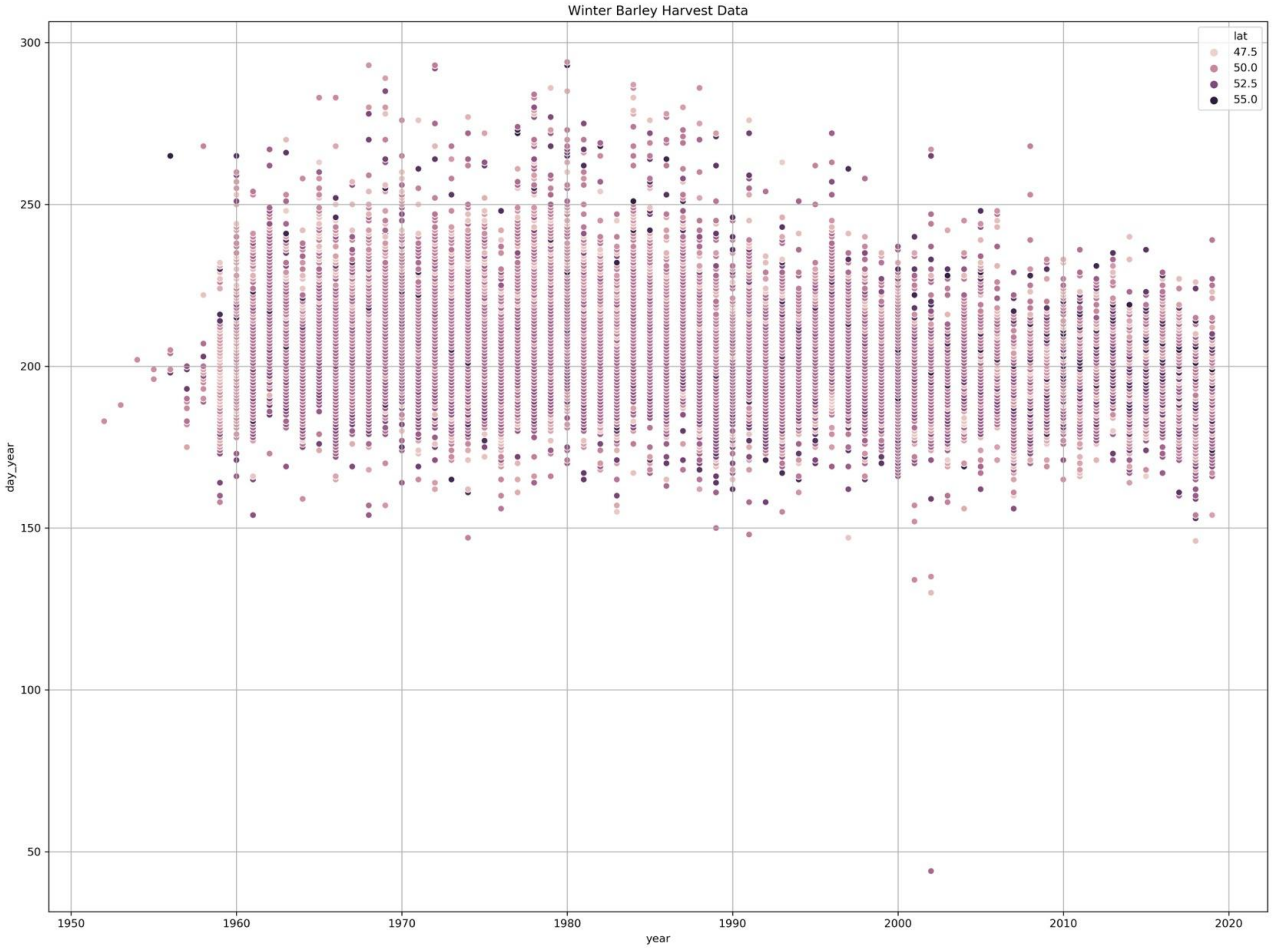
Barley phenological data

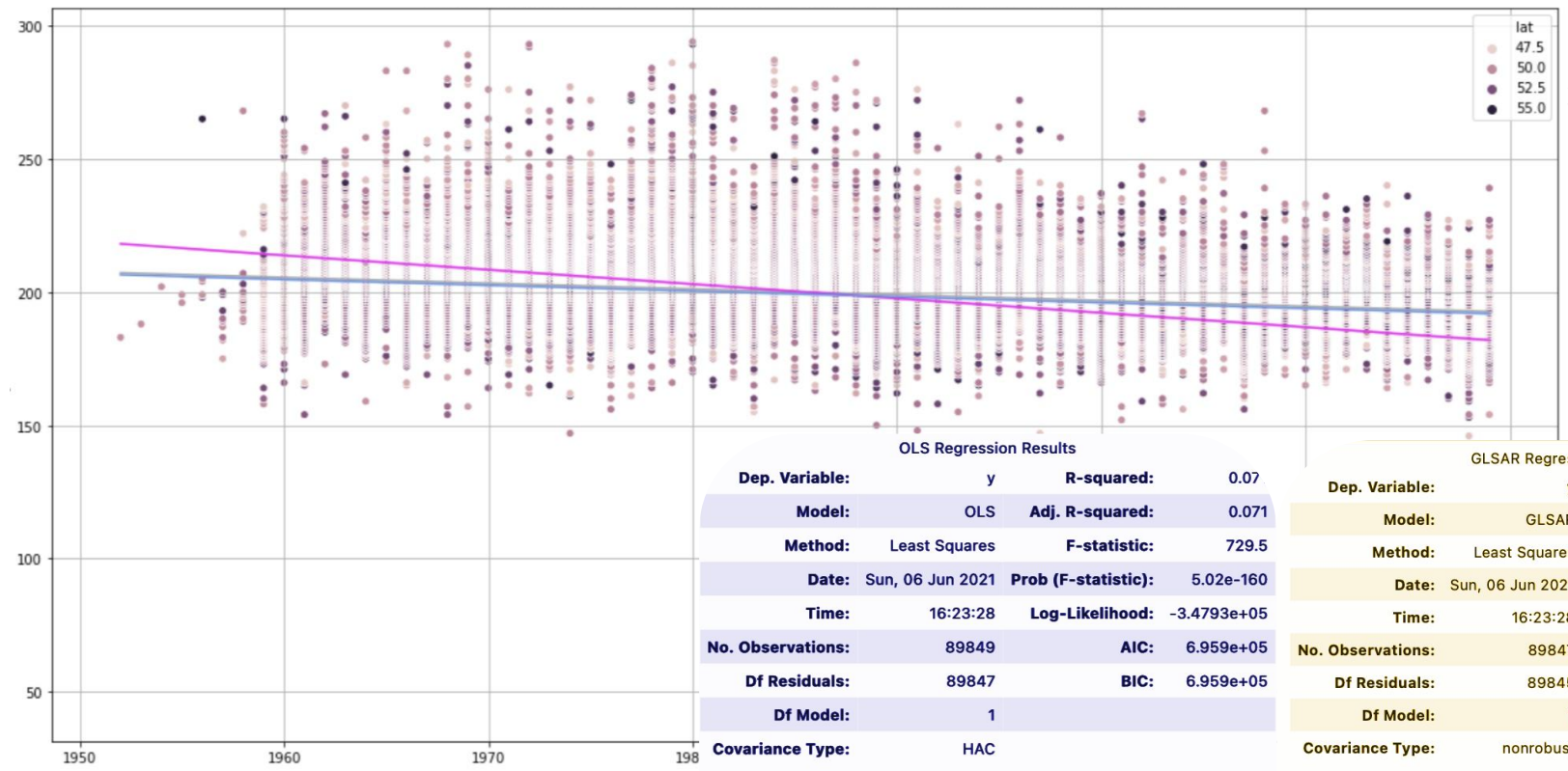
Only data on winter barley are available until 2019



Studying harvest dates for winter barley







X=year
y=first_harvest_day

Negative trend statistically significant

REGRESSION WITH OLS AND GLS

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.07
Model:	OLS	**Adj. R-squared:**	0.071
Method:	Least Squares	**F-statistic:**	729.5
Date:	Sun, 06 Jun 2021	**Prob (F-statistic):**	5.02e-160
Time:	16:23:28	**Log-Likelihood:**	-3.4793e+05
No. Observations:	89849	**AIC:**	6.959e+05
Df Residuals:	89847	**BIC:**	6.959e+05
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	630.0578	15.945	39.515	0.000	598.807	661.309
year	-0.2166	0.008	-27.009	0.000	-0.232	-0.201

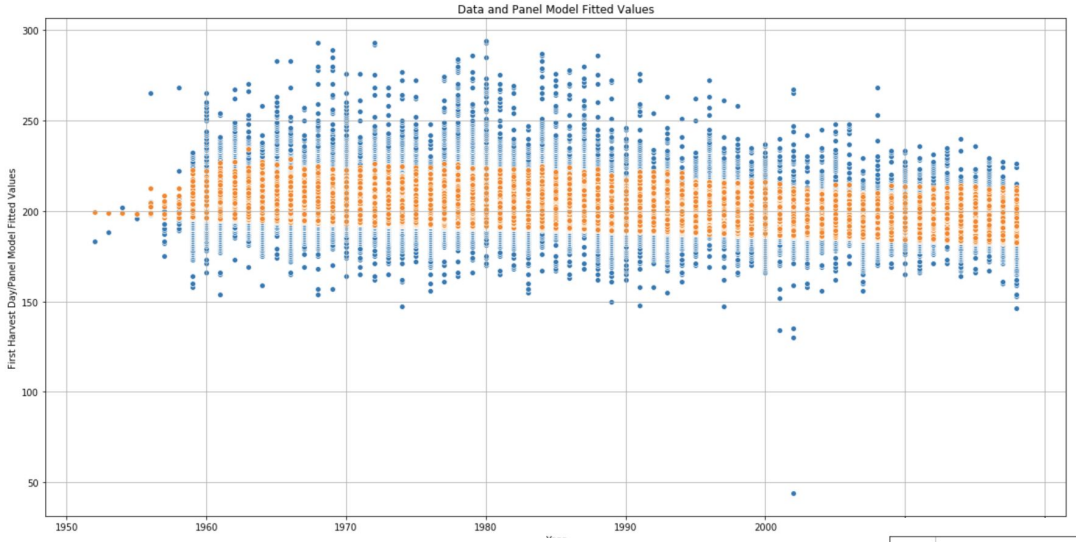
Omnibus:	16412.522	**Durbin-Watson:**	0.057
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	54473.754
Skew:	0.924	**Prob(JB):**	0.00
Kurtosis:	6.337	**Cond. No.**	2.65e+05

GLSAR Regression Results

Dep. Variable:	y	**R-squared:**	0.012
Model:	GLSAR	**Adj. R-squared:**	0.012
Method:	Least Squares	**F-statistic:**	1058.
Date:	Sun, 06 Jun 2021	**Prob (F-statistic):**	9.68e-231
Time:	16:23:28	**Log-Likelihood:**	-2.1743e+05
No. Observations:	89847	**AIC:**	4.349e+05
Df Residuals:	89845	**BIC:**	4.349e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1270.3682	32.910	38.601	0.000	1205.865	1334.871
year	-0.5391	0.017	-32.527	0.000	-0.572	-0.507

Omnibus:	249910.460	**Durbin-Watson:**	2.016
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9309535577.026
Skew:	-36.149	**Prob(JB):**	0.00
Kurtosis:	1578.292	**Cond. No.**	1.53e+05



PanelOLS Estimation Summary

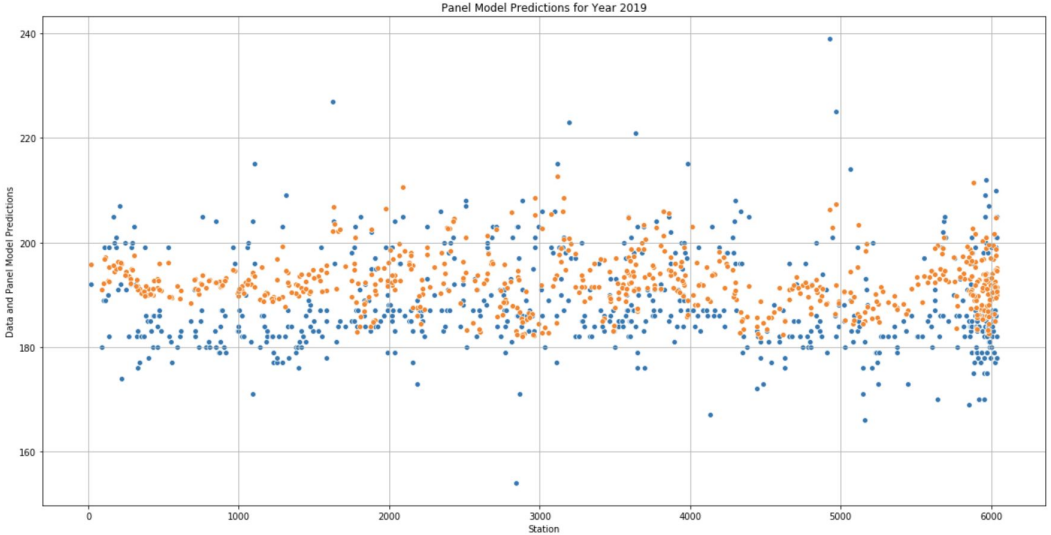
Dep. Variable:	day_year	R-squared:	0.2353
Estimator:	PanelOLS	R-squared (Between):	0.4264
No. Observations:	89198	R-squared (Within):	0.0690
Date:	Mon, Jun 07 2021	R-squared (Overall):	0.2353
Time:	16:45:27	Log-likelihood	-3.366e+05
Cov. Estimator:	Clustered		
		F-statistic:	6860.8
Entities:	5623	P-value	0.0000
Avg Obs:	15.863	Distribution:	F(4,89193)
Min Obs:	0.0000		
Max Obs:	60.000	F-statistic (robust):	1395.7
		P-value	0.0000
Time periods:	68	Distribution:	F(4,89193)
Avg Obs:	1311.7		
Min Obs:	0.0000		
Max Obs:	2275.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	480.09	9.1442	52.502	0.0000	462.17	498.01
lon	-1.2328	0.0421	-29.310	0.0000	-1.3152	-1.1504
lat	3.6511	0.0842	43.344	0.0000	3.4860	3.8162
alt	0.0474	0.0009	55.087	0.0000	0.0457	0.0491
year_dummy	-0.2339	0.0042	-55.661	0.0000	-0.2421	-0.2256

FORECASTING

The panel model is able to find a negative trend (coefficient for year_dummy is negative), but it can only reproduce in part the training data.



Thanks to

The teachers at Spiced Academy and my coursemates

The Agentur für Arbeit

PEP725 (Paneuropean Phenology Project)