

# Wine Project

Sergio Martinez Barajas      Sarah Marie Jennings

2025-11-30

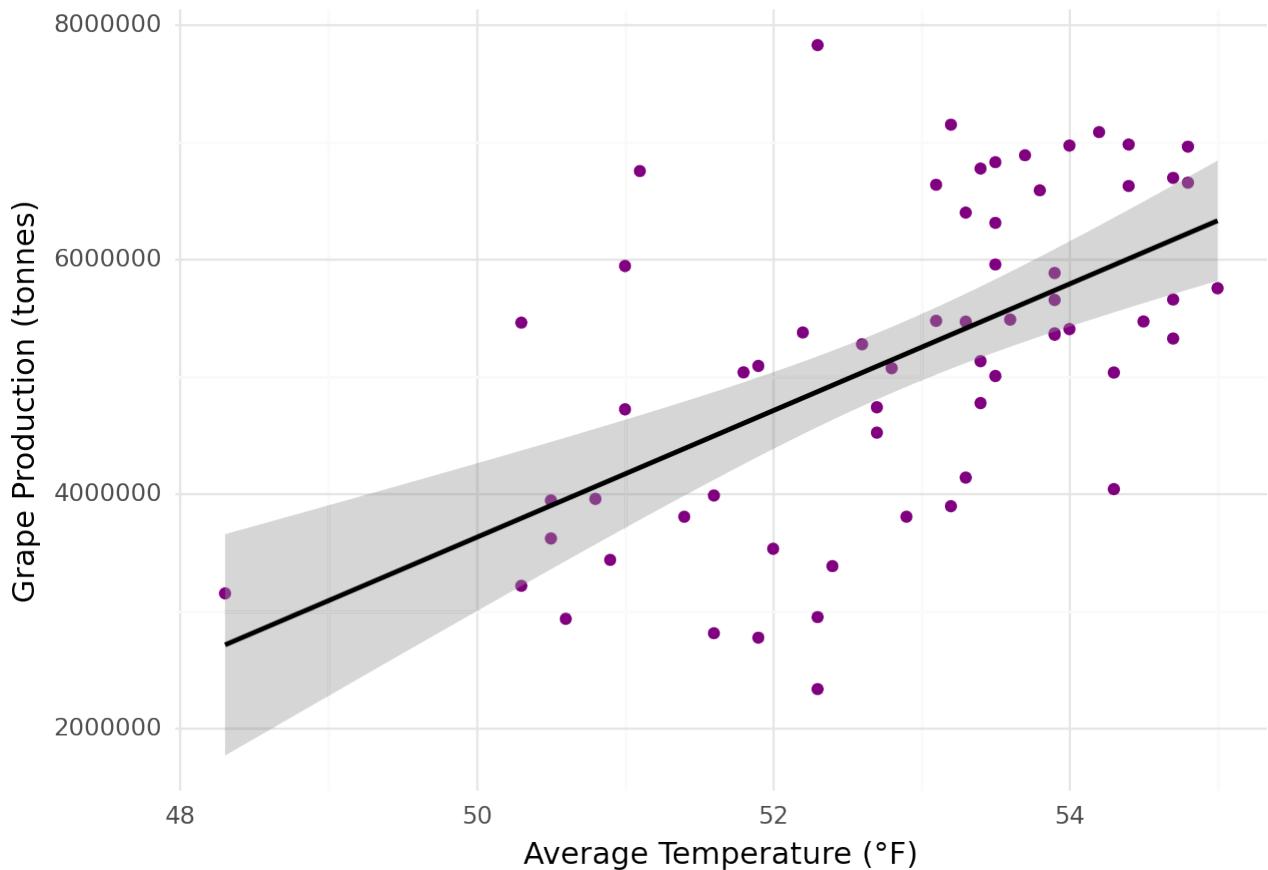
```
grapes = pd.read_csv("us-grape-production.csv")
temps = pd.read_csv("filtered-temp.csv")

grapes = grapes.rename(columns={
    "Grapes | 00000560 || Production | 005510 || tonnes": "Grape_Production_tonnes"
})

grapes = grapes[(grapes["Year"] >= 1961) & (grapes["Year"] <= 2023)]
temps = temps[(temps["Year"] >= 1961) & (temps["Year"] <= 2023)]

merged = pd.merge(grapes, temps, on="Year")
(
    ggplot(merged, aes(x="Average_Fahrenheit_Temperature", y="Grape_Production_tonnes")) +
    geom_point(color="purple") +
    geom_smooth(method="lm", color="black") +
    labs(
        title="U.S. Grape Production vs. Average Temperature (1961-2023)",
        x="Average Temperature (°F)",
        y="Grape Production (tonnes)"
    ) +
    theme_minimal()
)
```

## U.S. Grape Production vs. Average Temperature (1961-2023)



```
#Correlation analysis
corr = merged["Average_Fahrenheit_Temperature"].corr(merged["Grape_Production_tonnes"])
print("Correlation:", corr)
```

```
Correlation: 0.5712335660023041
```

what the correlation means is that there tends to be positive correlation between average yearly temperature and grape yield, but it's not very strong.

```
#linear regression
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np
```

```

X = merged["Average_Fahrenheit_Temperature"].values.reshape(-1, 1)
Y = merged["Grape_Production_tonnes"].values

X_train, X_test, Y_train, Y_test = train_test_split( X, Y, test_size = 0.2, shuffle = True, random_state = 42)

model = LinearRegression()
model.fit(X_train, Y_train)

Y_pred = model.predict(X_test)

mae = mean_absolute_error(Y_test, Y_pred)
rmse = np.sqrt(mean_squared_error(Y_test, Y_pred))
r2 = r2_score(Y_test, Y_pred)

print("MAE: ", mae)
print("RMSE: ", rmse)
print("R^2: ", r2)

```

```

MAE:  1019018.7867519022
RMSE:  1162452.7731372928
R^2:  0.32537166473166856

```

slope means that for every 1 F increase we get 540k tonnes increase yield but propably because temperature is average across all US, grape production is national total, there are other possible factors at work  $R^2 = 0.326$  means that temperture explains 32% of the variation in grape yields, which means that a whole 2/3rds of variablility is caused by something else.

Both temp and yield increase overtime so this might make false impressions. It's not that more heat = more grapes, it could be more time = more grapes and more heat. So if we detrend the data to see if the correlation still exists.

```

merged["Year_centered"] = merged["Year"] - merged["Year"].mean()

# detrend
from sklearn.linear_model import LinearRegression

# detrend temperature
m1 = LinearRegression().fit(merged[["Year_centered"]], merged["Average_Fahrenheit_Temperature"])
merged["Temp_detrended"] = merged["Average_Fahrenheit_Temperature"] - m1.predict(merged[["Year_centered"]])

# detrend yield

```

```

m2 = LinearRegression().fit(merged[["Year_centered"]], merged["Grape_Production_tonnes"])
merged["Yield_detrended"] = merged["Grape_Production_tonnes"] - m2.predict(merged[["Year_centered"]])

merged[["Temp_detrended"]].corr(merged["Yield_detrended"])

```

-0.005950884596132765

After detrending the data we see that correlation is basically 0. So temp and grape yield have almost nothing to do with each other.

```

#####
##### From here on out we're only using Cal Data #####
#####

grapes = pd.read_csv("California_Wine_Production_1980_2020.csv")
temp = pd.read_csv("California-avg-temp-1980-2021.csv")
rain = pd.read_csv("California-rain-1980-2021.csv")

# trim to overlapping years
grapes = grapes[(grapes["Year"] >= 1980) & (grapes["Year"] <= 2020)]
temp = temp[(temp["Year"] >= 1980) & (temp["Year"] <= 2020)]
rain = rain[(rain["Year"] >= 1980) & (rain["Year"] <= 2020)]

# merge
merged = grapes.merge(temp, on="Year").merge(rain, on="Year")

merged = merged.rename(columns={
    "Temp": "AvgTemp_F",
    "rain_in": "Rain_in"
})
merged[["Yield(Unit/Acre)", "AvgTemp_F", "Rain_in"]].corr()

```

	Yield(Unit/Acre)	AvgTemp_F	Rain_in
Yield(Unit/Acre)	1.000000	0.022516	0.002424
AvgTemp_F	0.022516	1.000000	-0.267420
Rain_in	0.002424	-0.267420	1.000000

What this means is that temp and precipitation have almost no impact on yield. Although let it be known that in our lit review it did say that precipitation almost never has an impact on

yields, so that's at least 1 thing proven. Temp and rain have a negative correlation, when it rains it's colder and when hot there's less rain, but it's really weak

```
county_stats = merged.groupby("County").agg({
    "Yield(Unit/Acre)": "mean",
    "HarvestedAcres": "mean",
    "AvgTemp_F": "mean",
    "Rain_in": "mean"
}).sort_values(by="HarvestedAcres", ascending=False)

print(county_stats)
```

County	Yield(Unit/Acre)	HarvestedAcres	AvgTemp_F	Rain_in
SanJoaquin	6.849268	70829.780488	58.717073	22.571463
Fresno	10.505854	61157.951220	58.717073	22.571463
Madera	9.824634	46011.853659	58.717073	22.571463
Sonoma	3.847073	42911.219512	58.717073	22.571463
SanLuisObisp	4.100000	36600.000000	57.500000	20.630000
Napa	3.647073	35201.341463	58.717073	22.571463
Kern	9.046341	35094.560976	58.717073	22.571463
Monterey	4.161389	35093.777778	58.830556	22.370000
SanLuisObispo	4.308889	22972.861111	58.766667	22.243333
Sacramento	7.032683	18104.365854	58.717073	22.571463
Tulare	11.260244	17996.536585	58.717073	22.571463
Stanislaus	8.937619	15855.428571	58.428571	23.613333
SantaBarbara	3.540556	14581.805556	58.730556	22.233333
Mendocino	4.051282	13917.717949	58.746154	22.495897
Merced	9.039024	13003.463415	58.717073	22.571463
Yolo	6.908537	7654.000000	58.717073	22.571463
Lake	4.030000	5392.731707	58.717073	22.571463
SanBenito	3.976098	3223.390244	58.717073	22.571463
Amador	3.440256	2888.820513	58.741026	22.727949
Solano	4.739756	2855.853659	58.717073	22.571463
Kings	11.305610	2615.341463	58.717073	22.571463
Riverside	3.841951	2266.341463	58.717073	22.571463
Alameda	3.678049	2124.097561	58.717073	22.571463
Colusa	8.633000	1896.700000	59.500000	21.298000
SanBernardino	2.461951	1791.341463	58.717073	22.571463
ContraCosta	3.699091	1616.818182	59.018182	24.374545
SantaClara	3.150833	1615.416667	58.711111	21.843889
ElDorado	3.022727	1458.424242	58.806061	23.114848

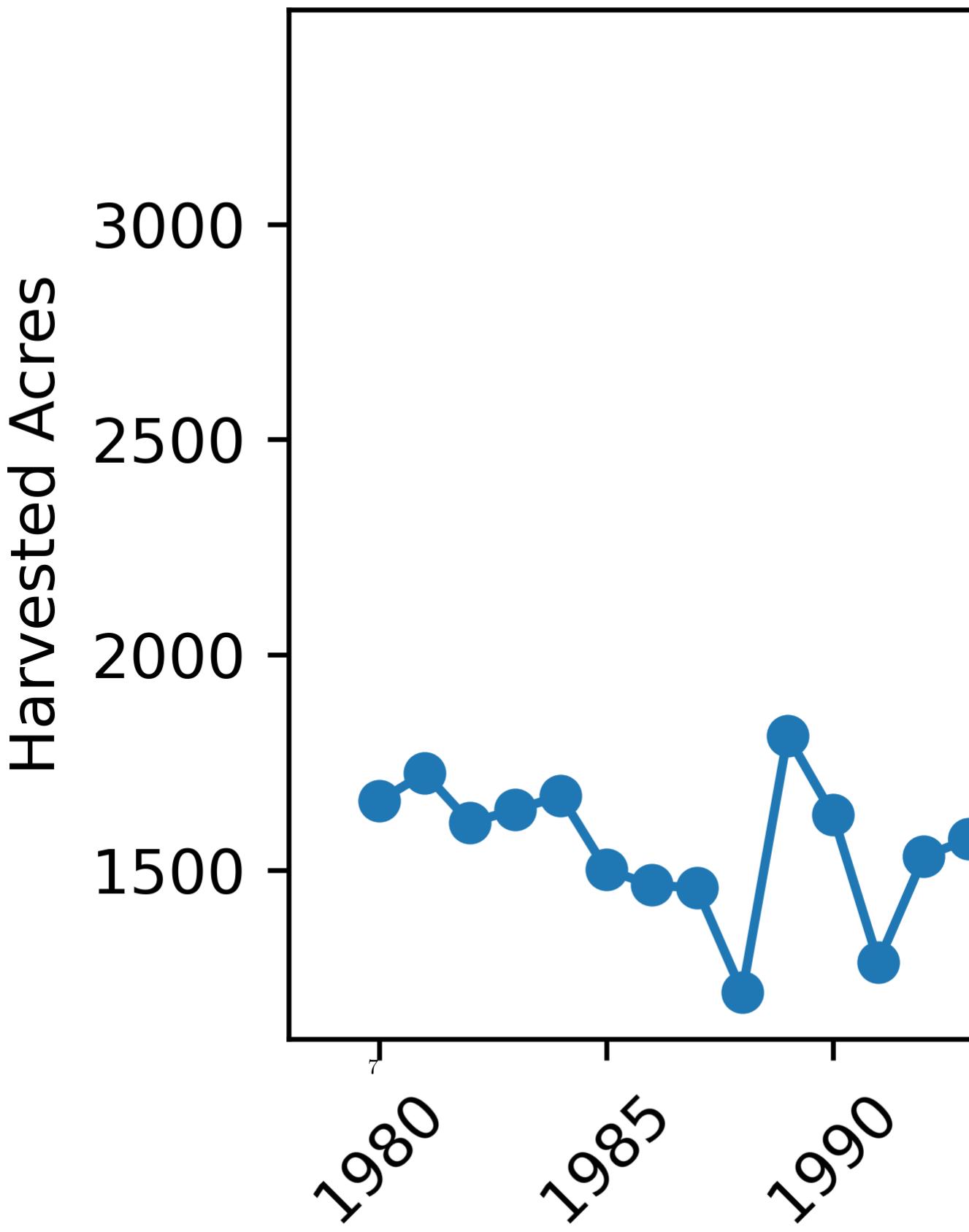
Glenn	6.750000	875.000000	57.750000	29.885000
Calaveras	2.694000	443.200000	58.737500	22.665750
SantaCruz	2.090833	422.305556	58.791667	22.632222
SanDiego	2.550244	416.097561	58.717073	22.571463
Nevada	4.037586	309.533333	58.744737	22.401316
Placer	2.457500	190.392857	58.864286	21.974643
Yuba	2.008571	172.500000	58.607143	22.031429
Shasta	2.358333	167.500000	59.425000	20.604167
Marin	1.485455	154.409091	59.027273	21.613182
Tehama	4.250000	147.000000	59.700000	22.506667
SanMateo	NaN	101.962963	59.014286	22.725000
Trinity	2.130000	94.687500	58.910000	22.110000
Mariposa	1.264000	80.400000	58.950000	21.981000
Mono	13.755000	14.250000	58.250000	22.597500

So right here we're looking at the counties that harvest the most winegrapes. Doesn't really tell us anything

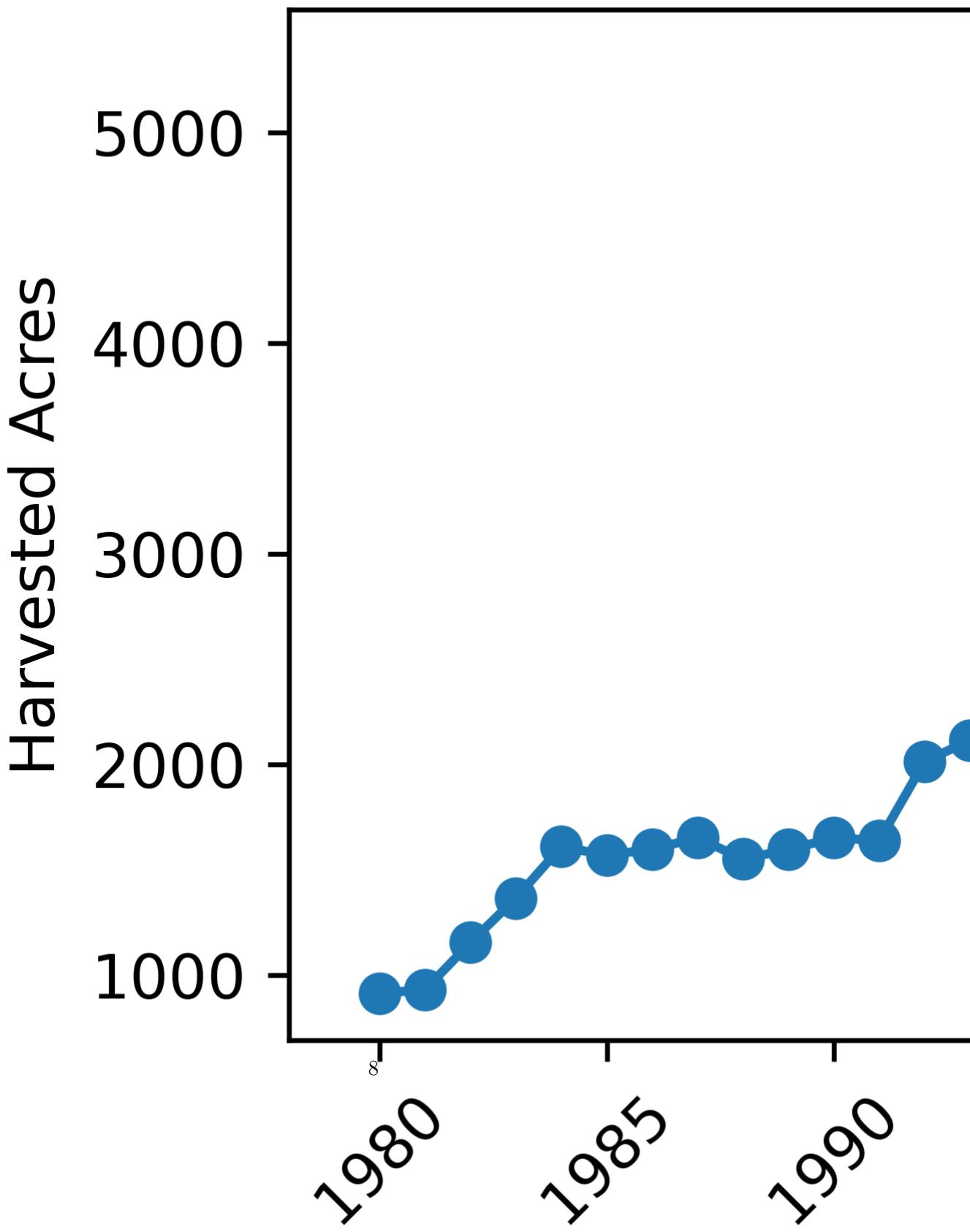
```
for county in merged["County"].unique():
    sub = merged[merged["County"] == county]

    if len(sub) > 1: # must have multiple years to plot
        plt.figure()
        plt.plot(sub["Year"], sub["HarvestedAcres"], marker="o")
        plt.title(f"Harvested Acres Over Time - {county}")
        plt.xlabel("Year")
        plt.ylabel("Harvested Acres")
        plt.xticks(rotation=45)
        plt.tight_layout()
        plt.show()
```

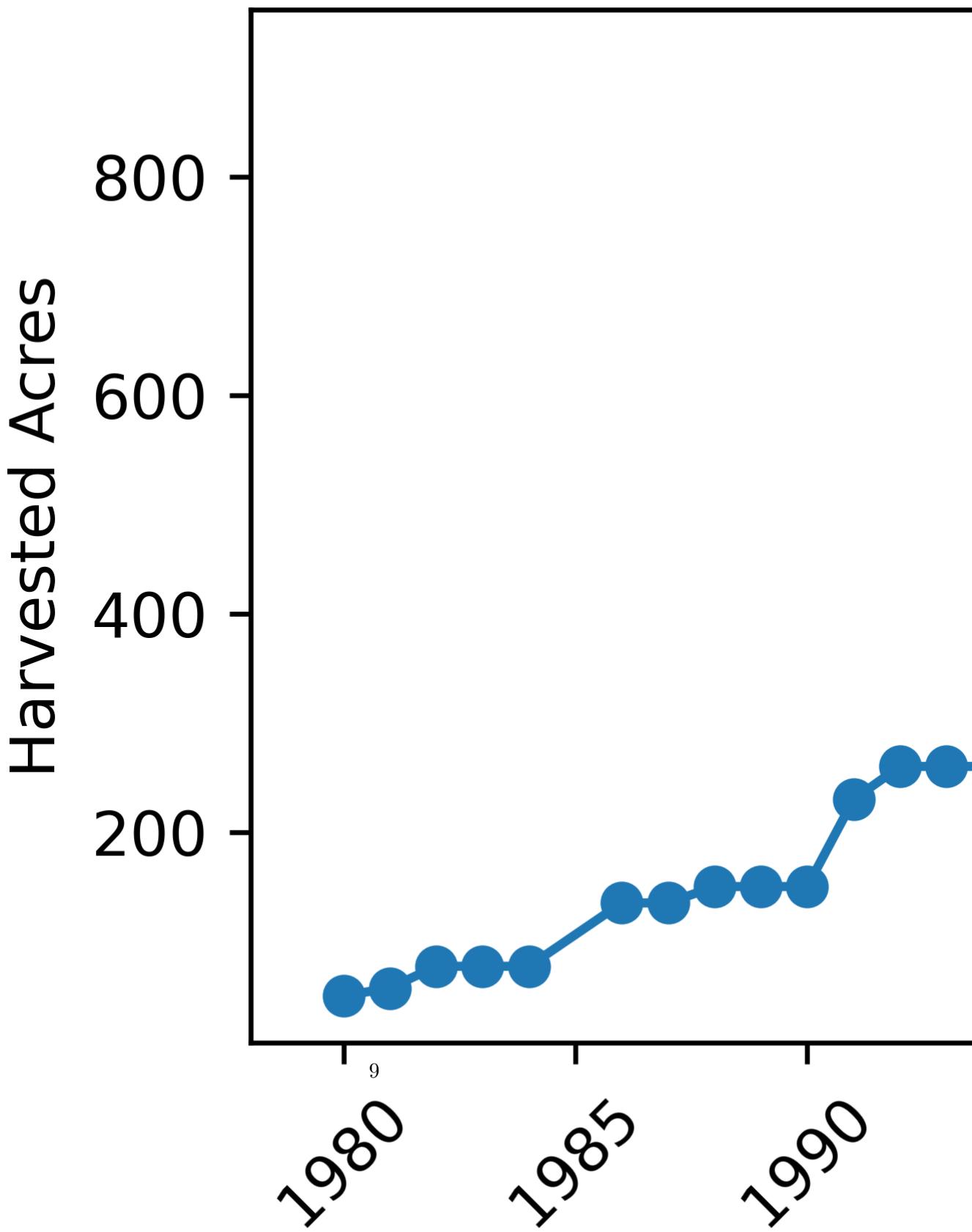
Harvested Acres



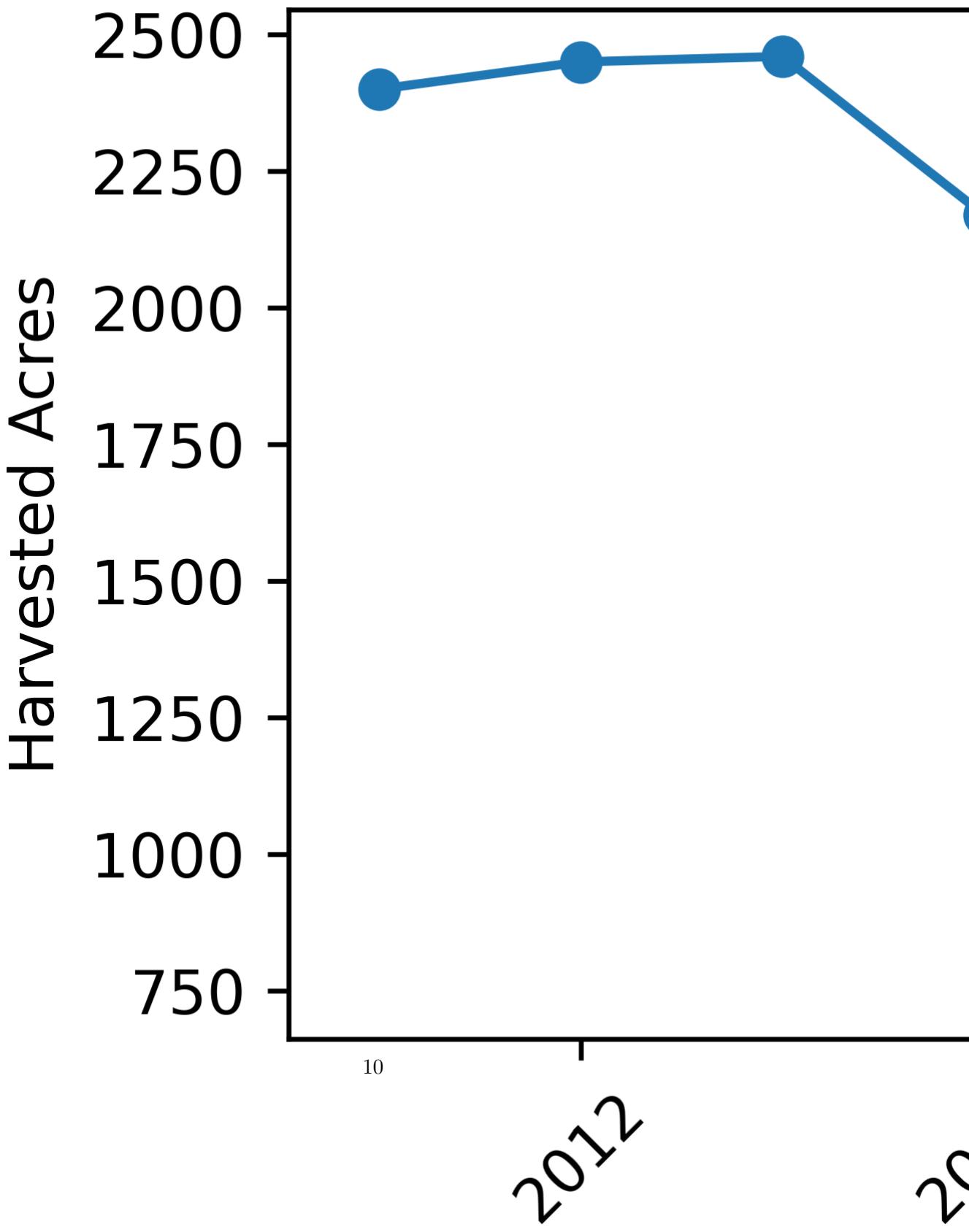
Harvested



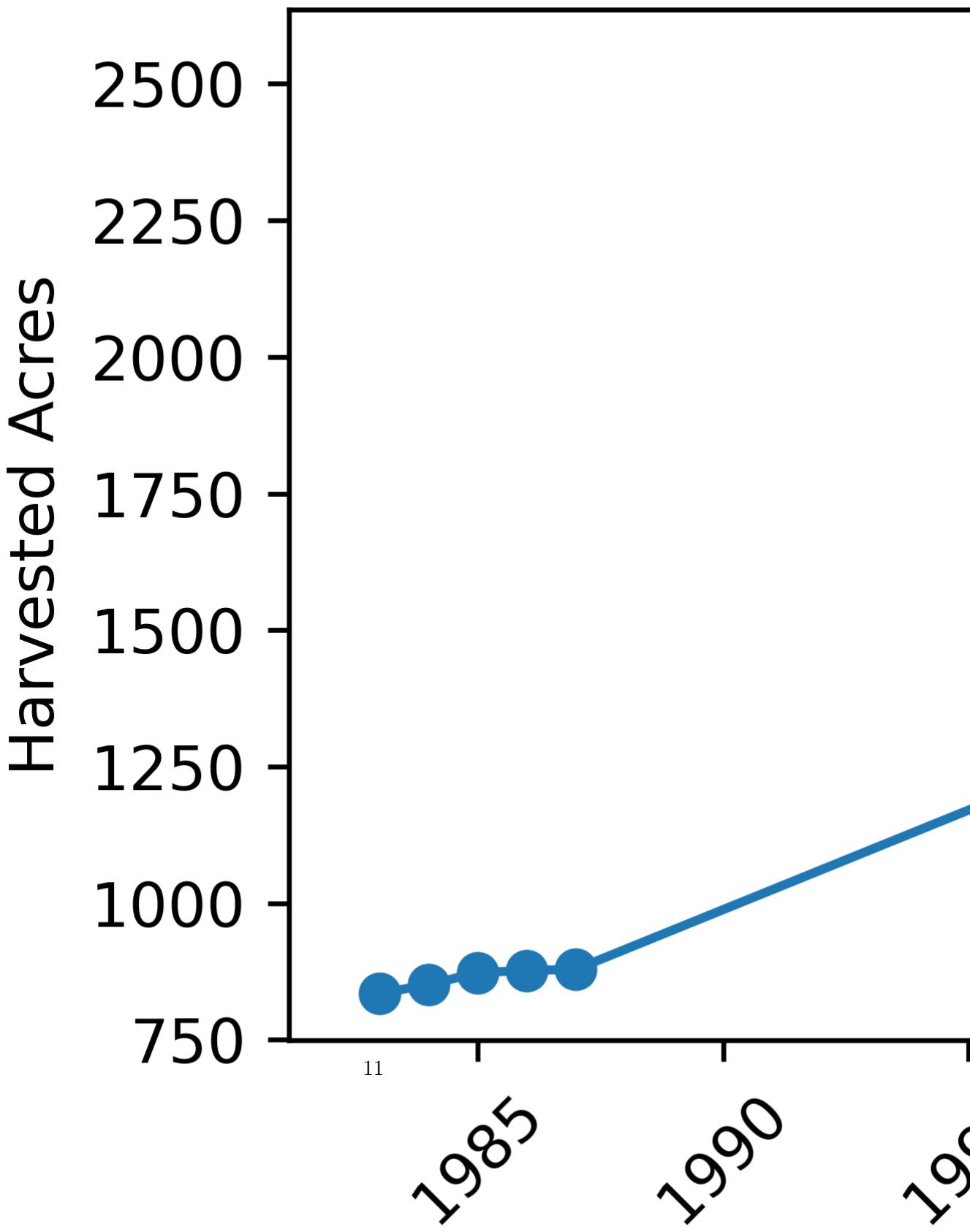
# Harvested Acres



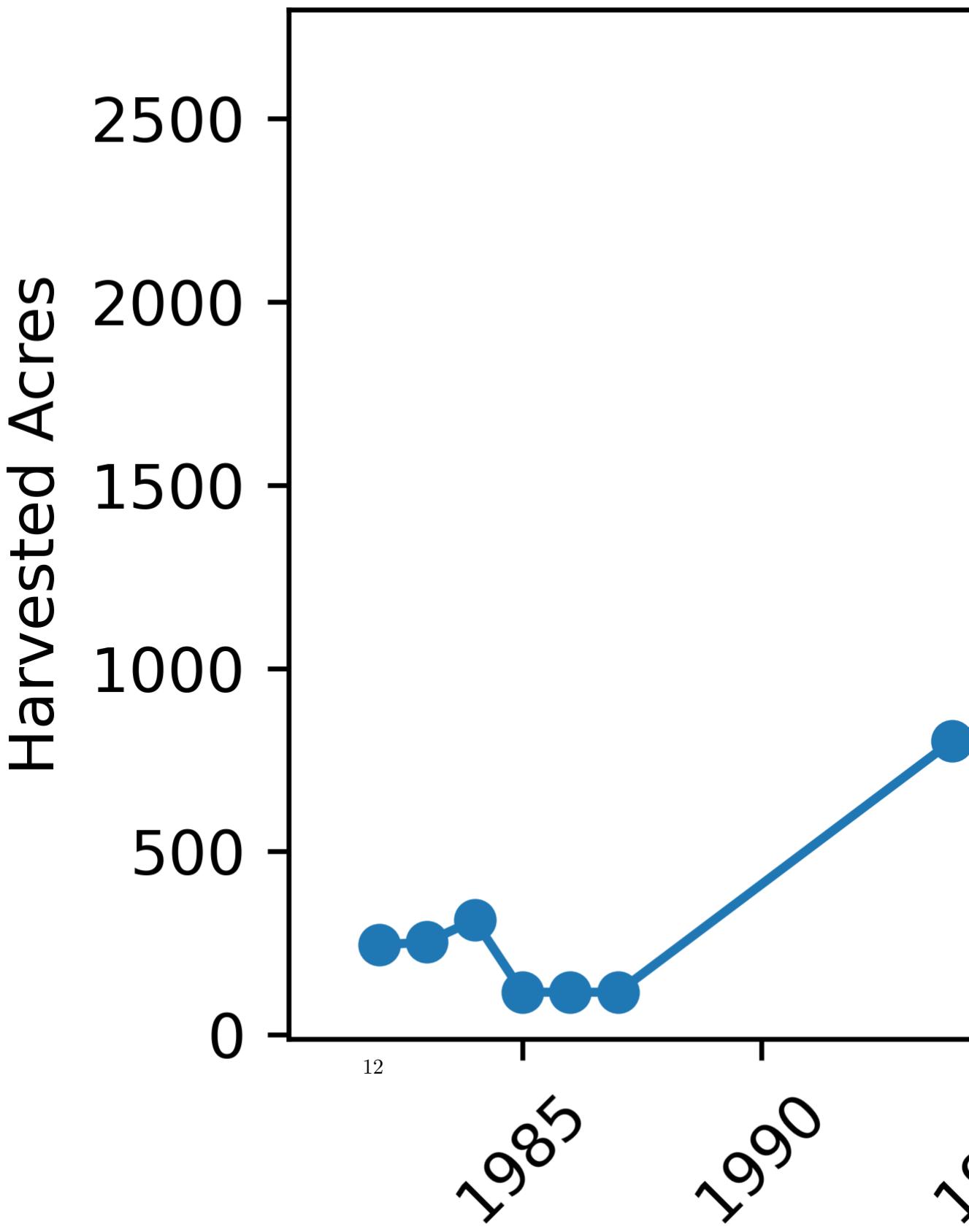
Harvested



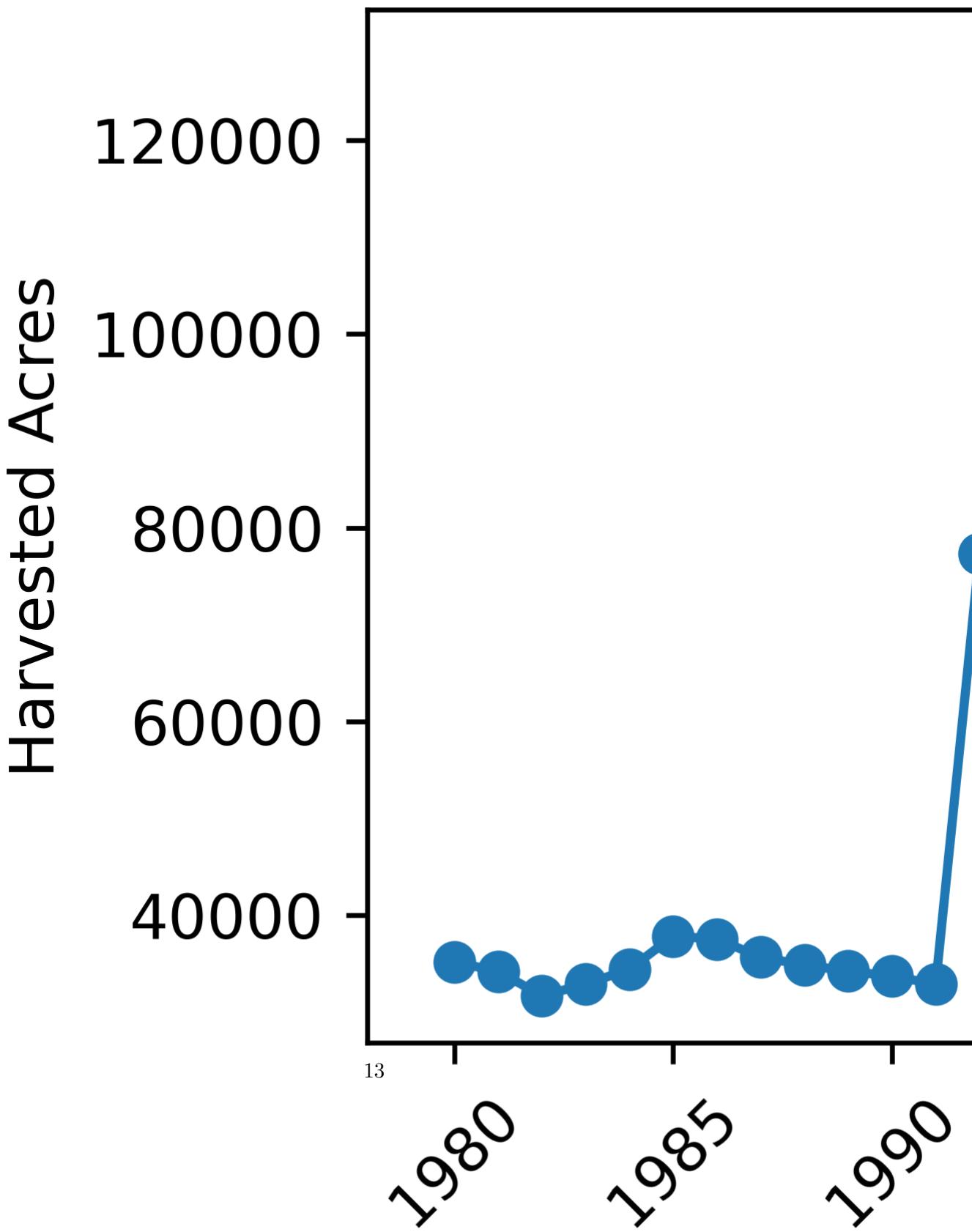
# Harvested Acres



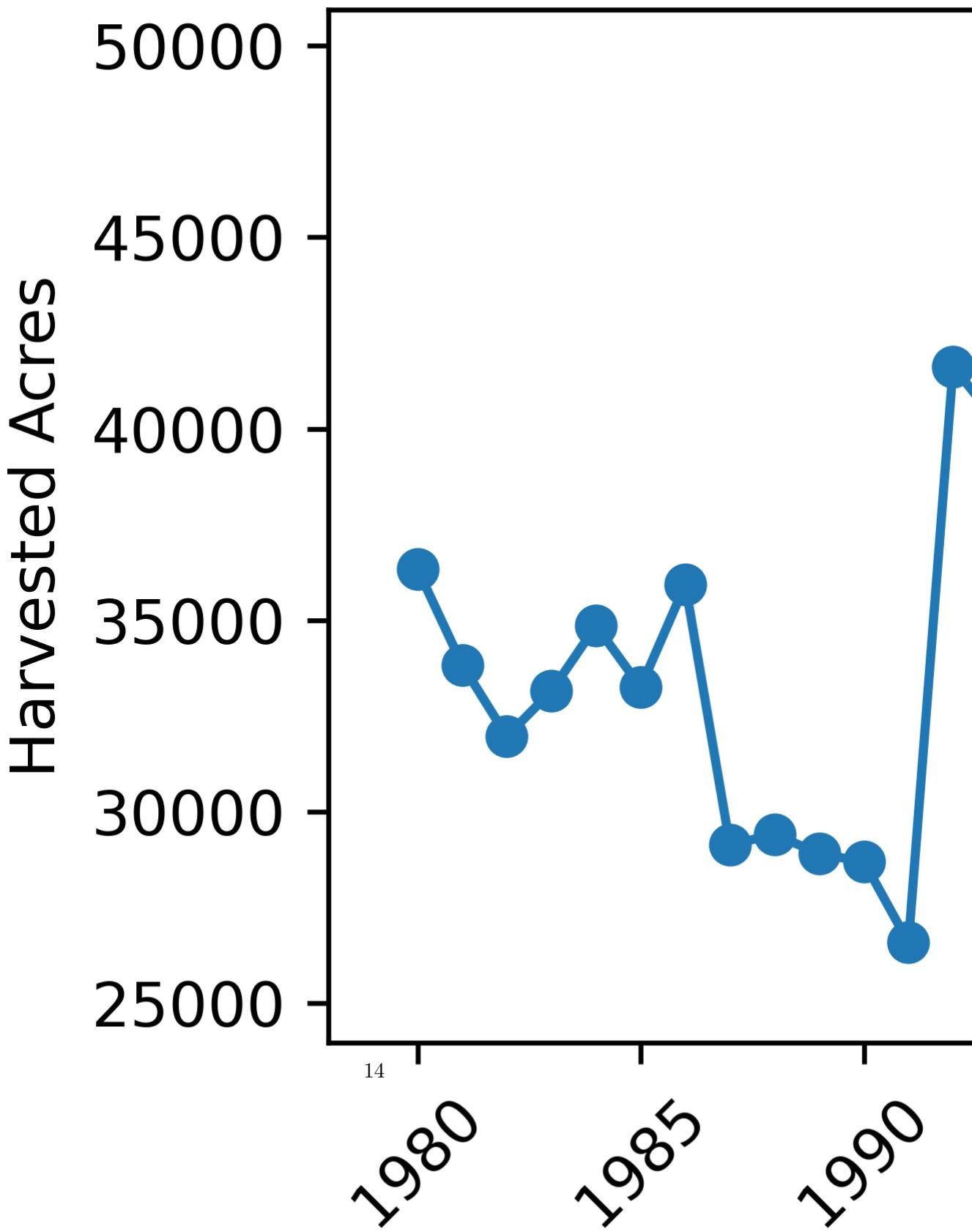
Harvested Acres



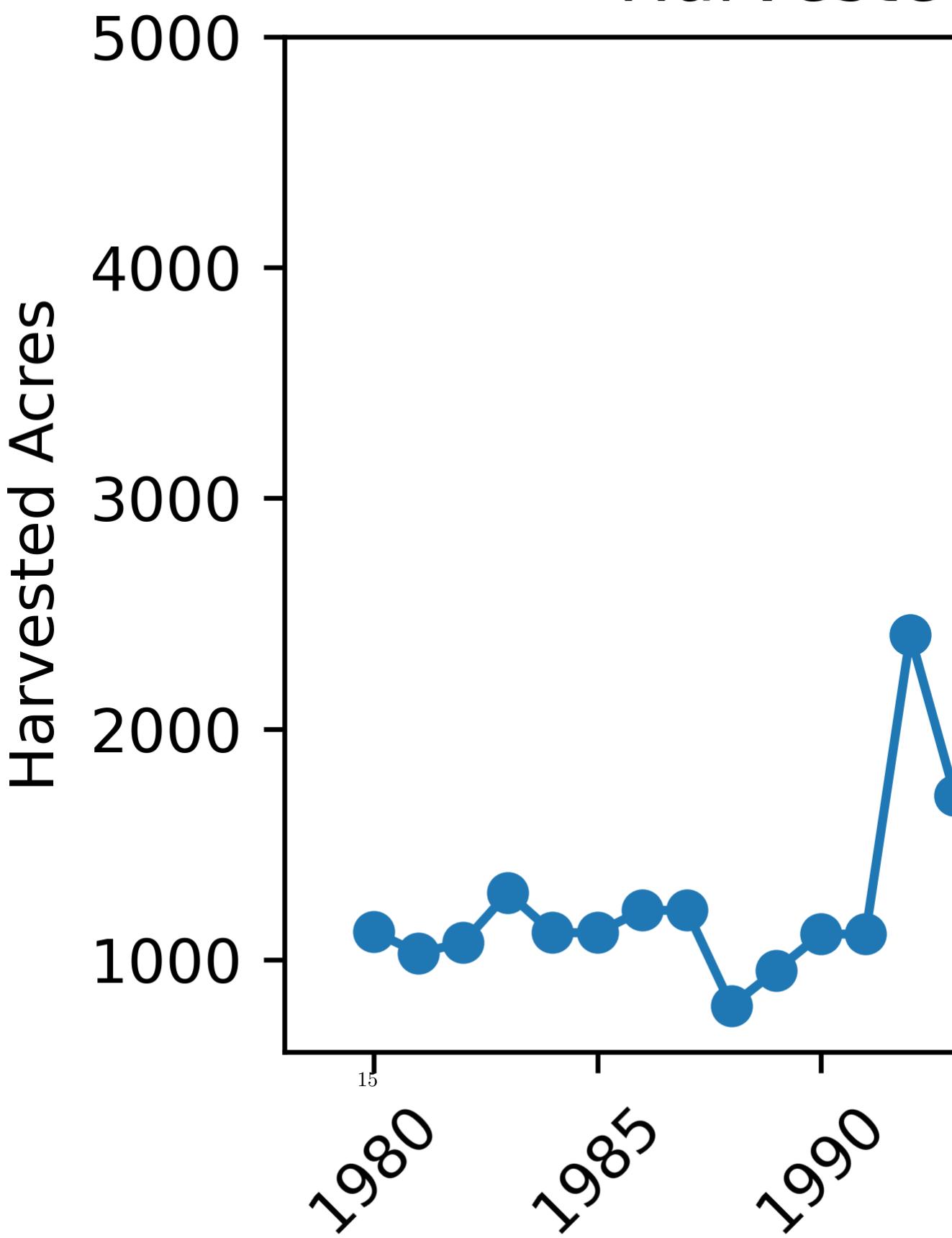
# Harvested Acres



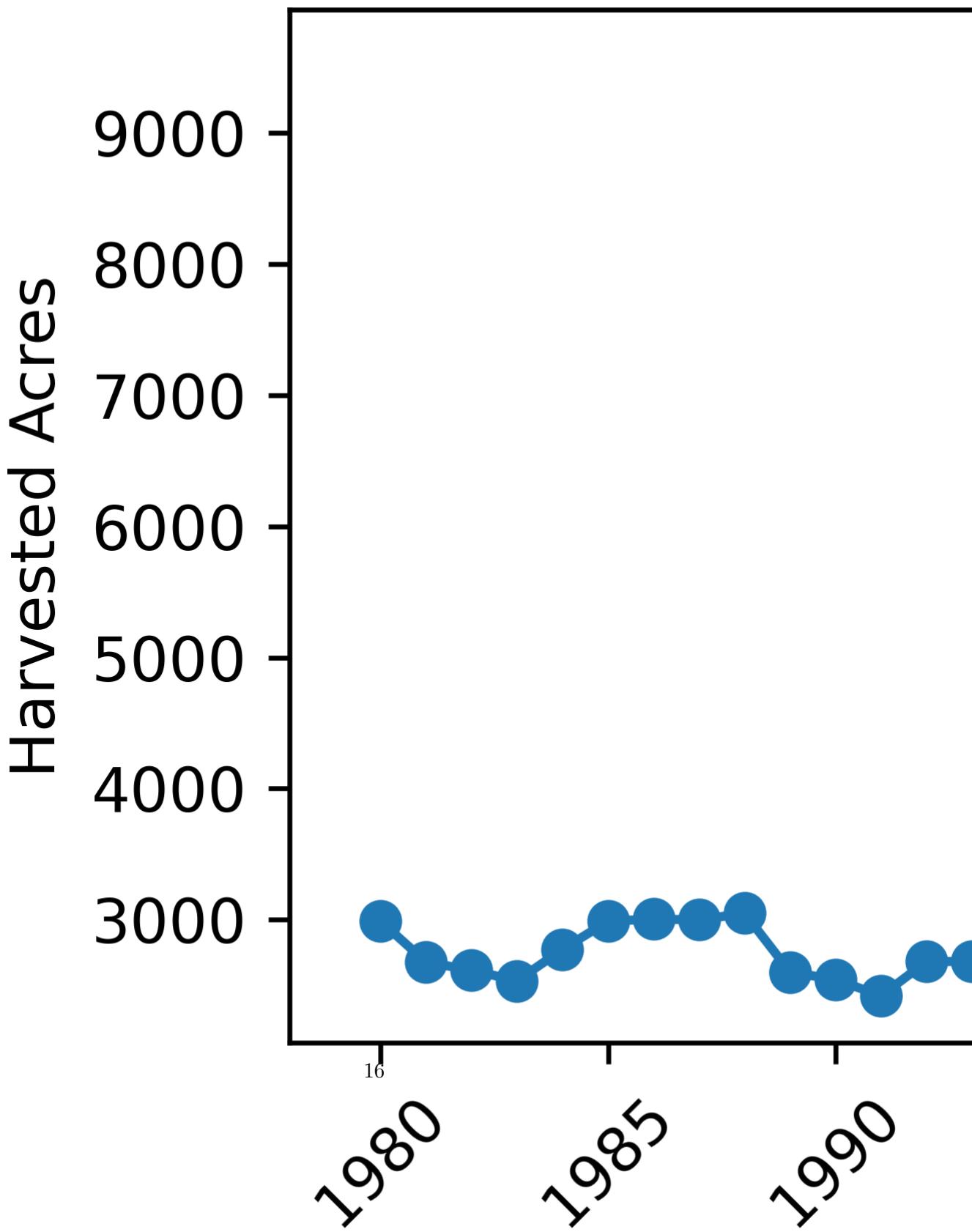
# Harvested Acres



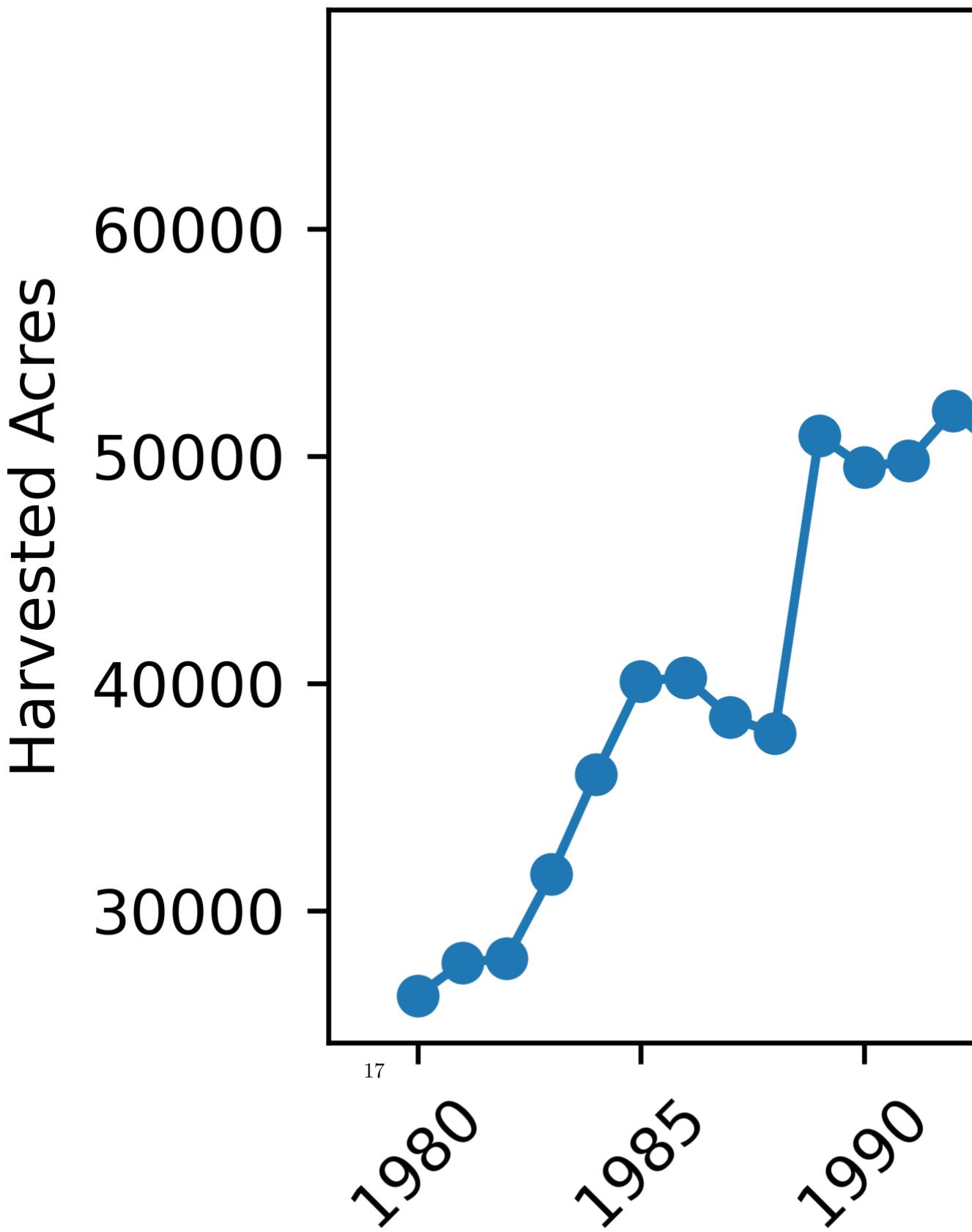
Harvested



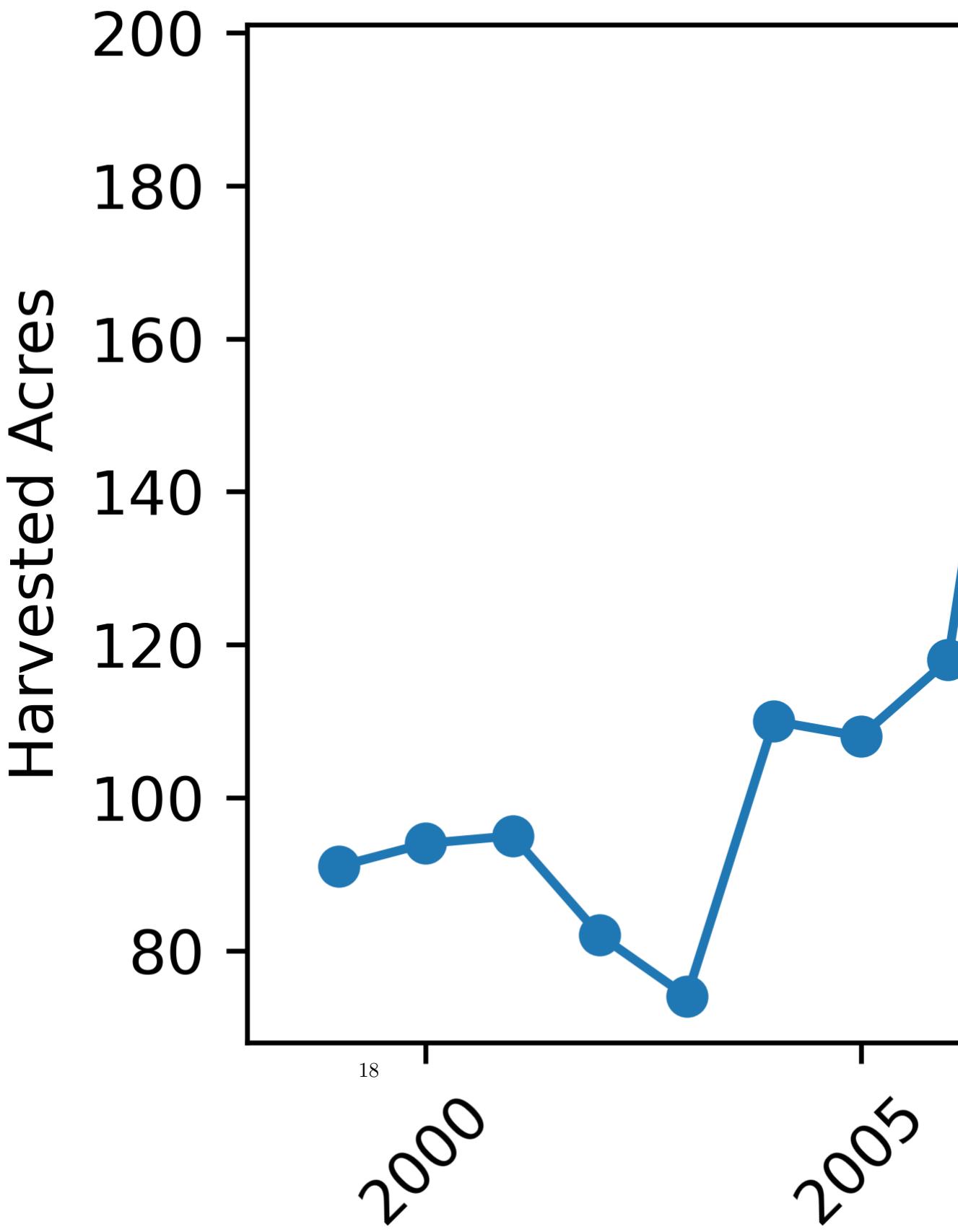
# Harvested Acres



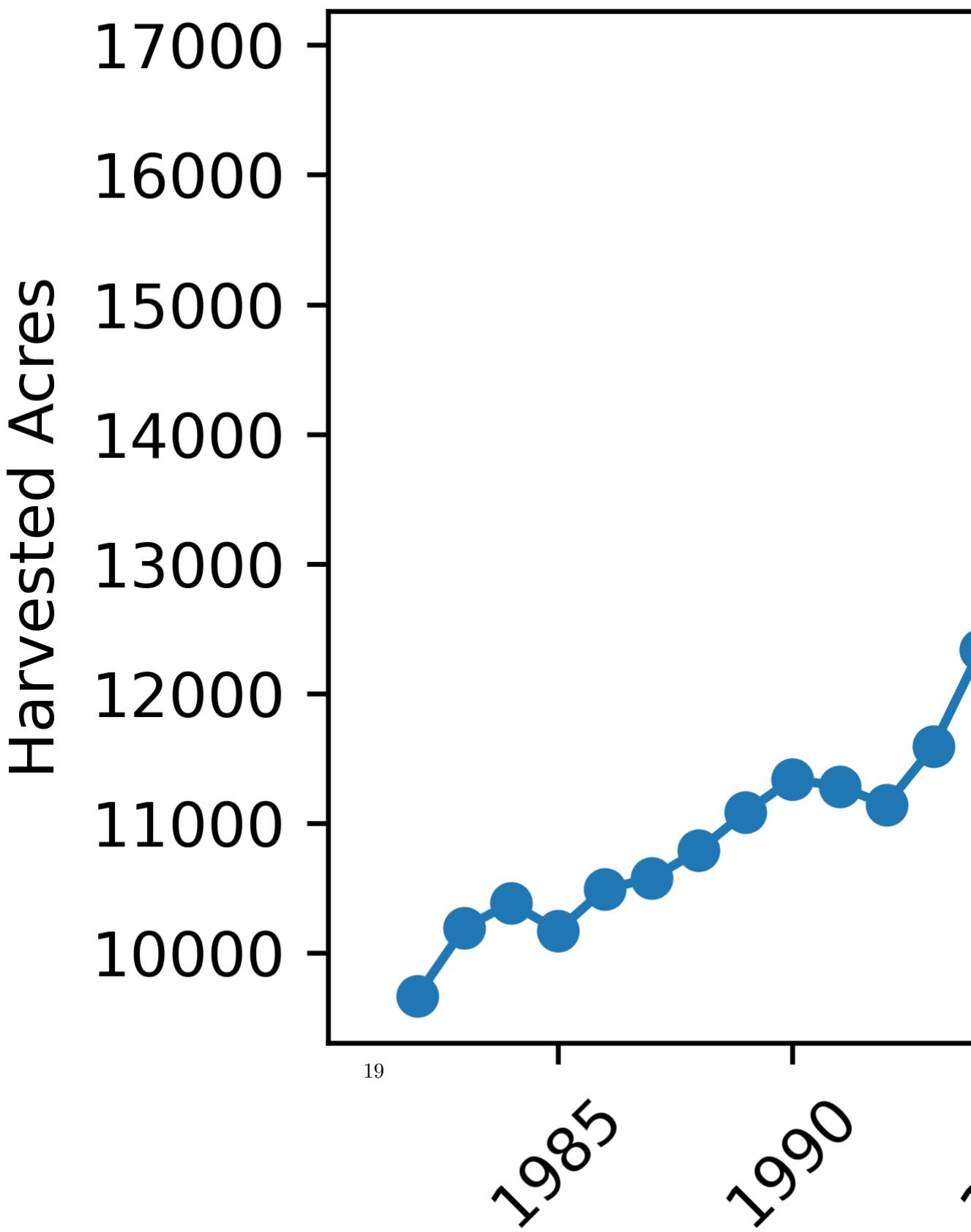
# Harvested Acres



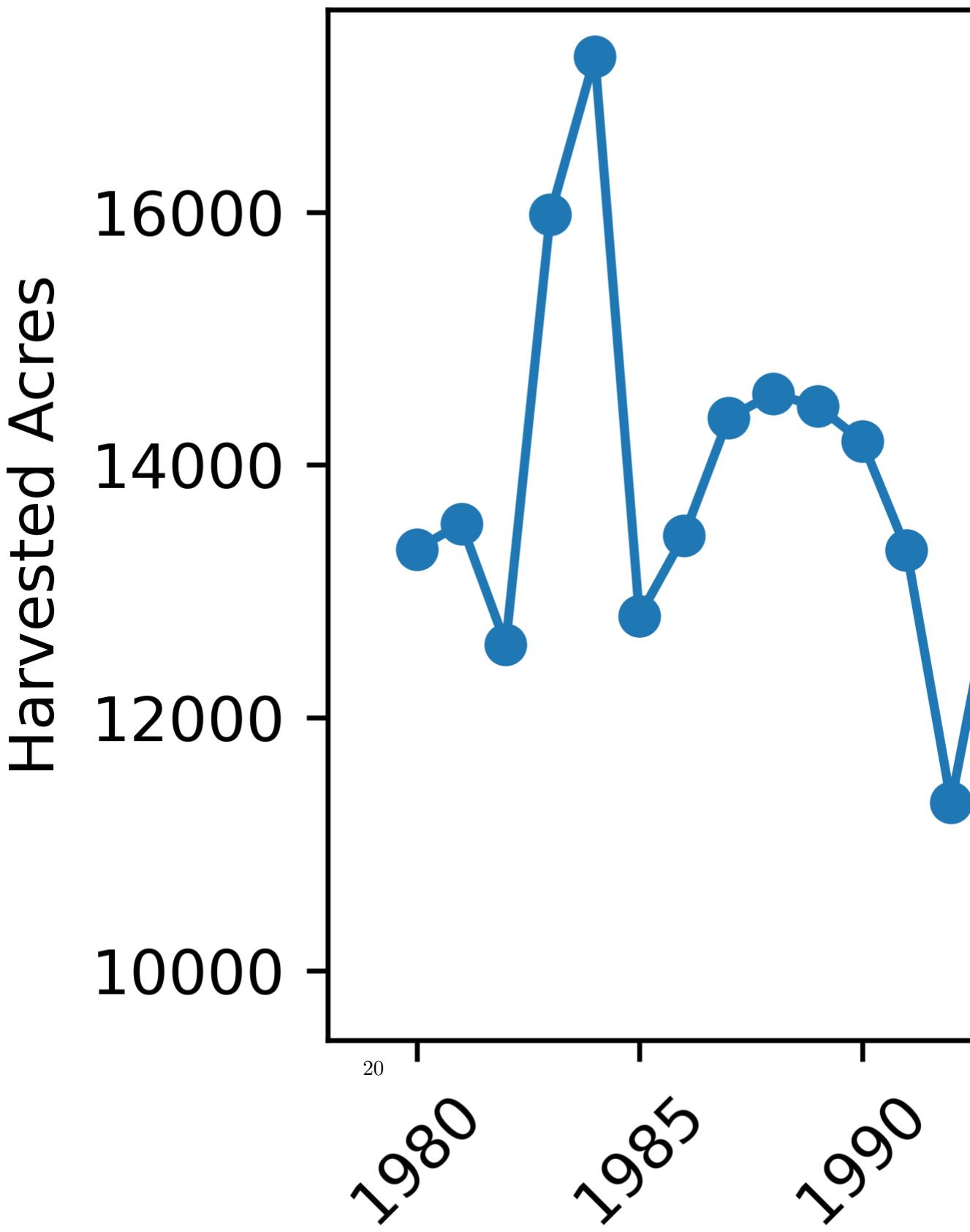
Harvested



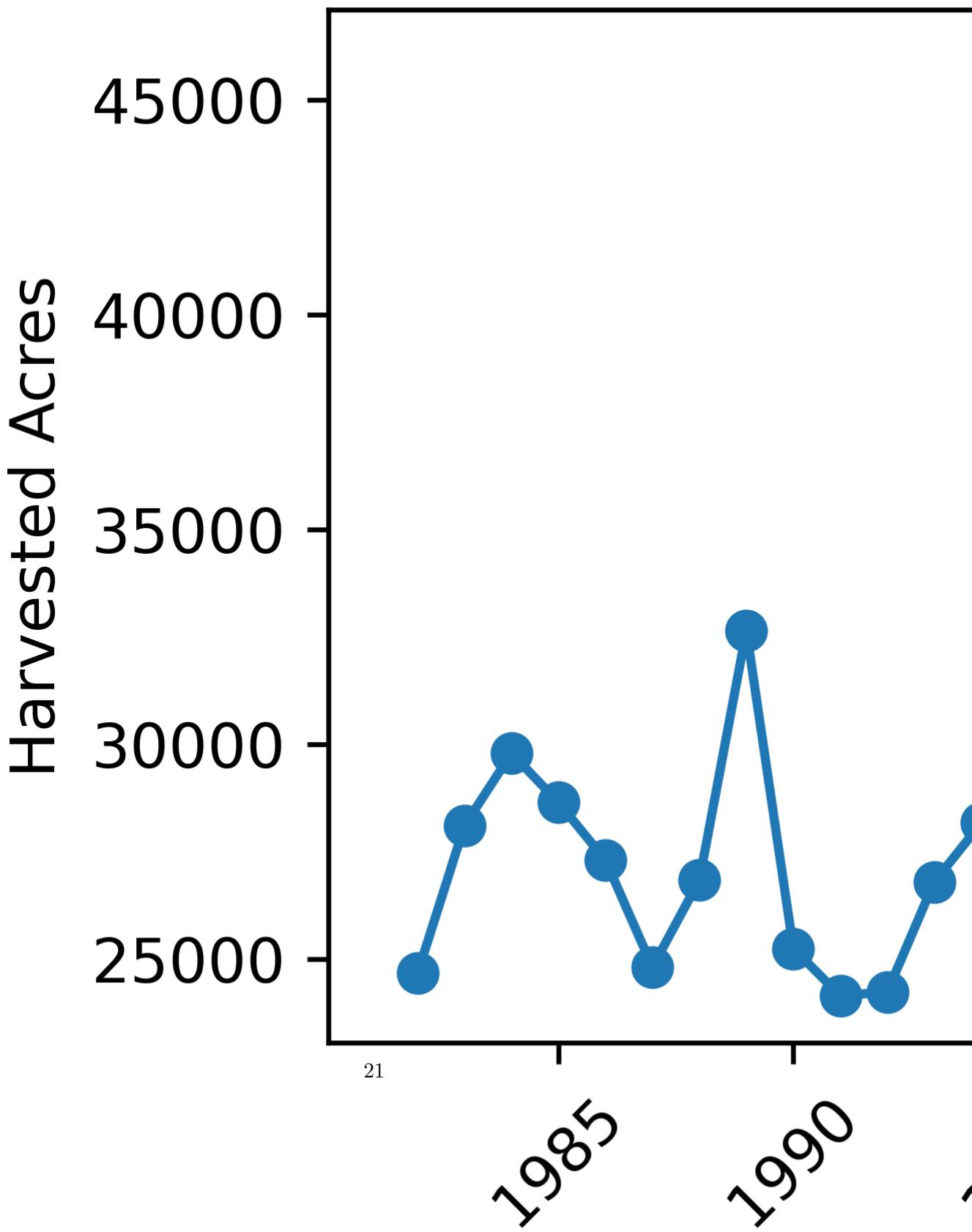
# Harvested Acres



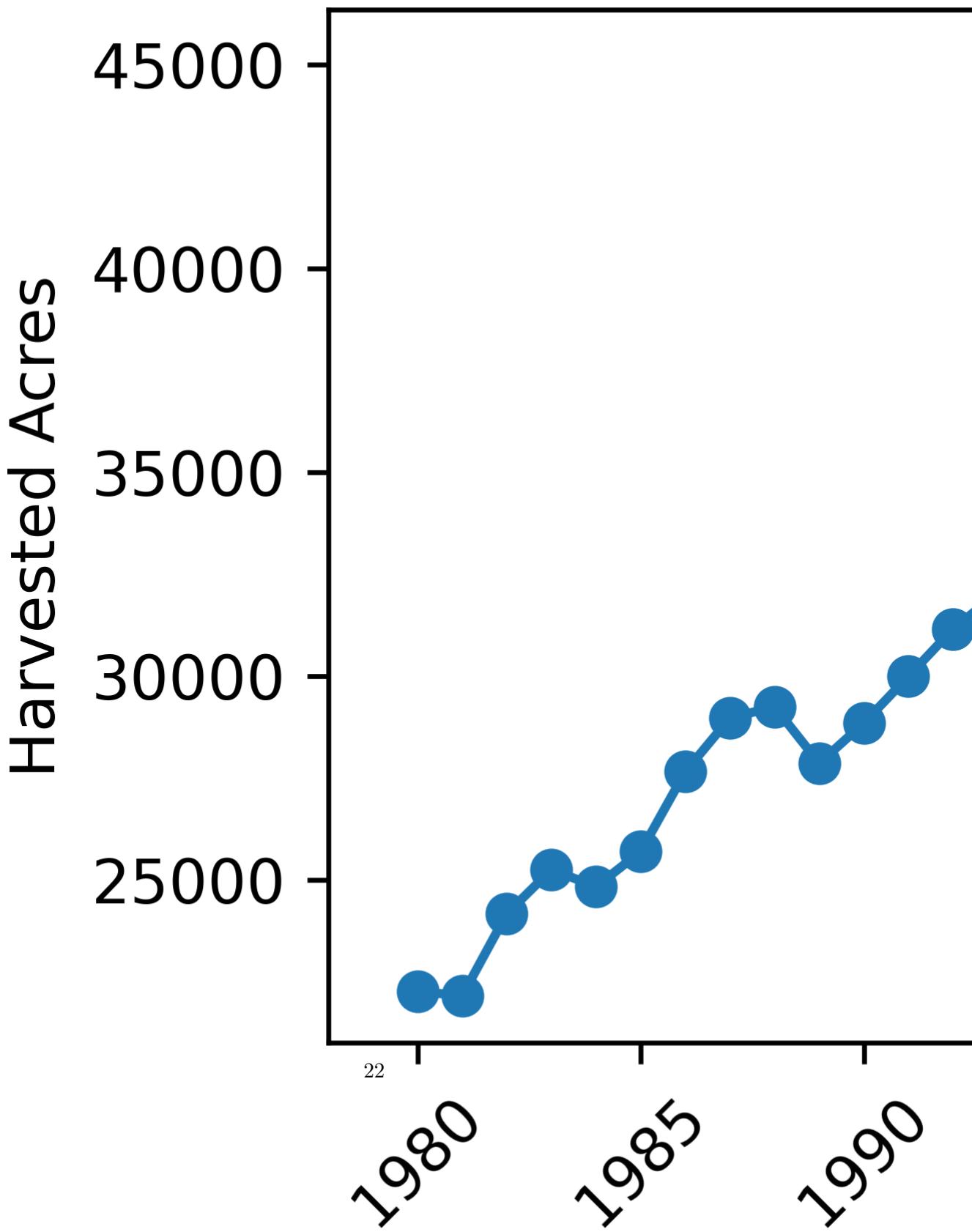
Harvested



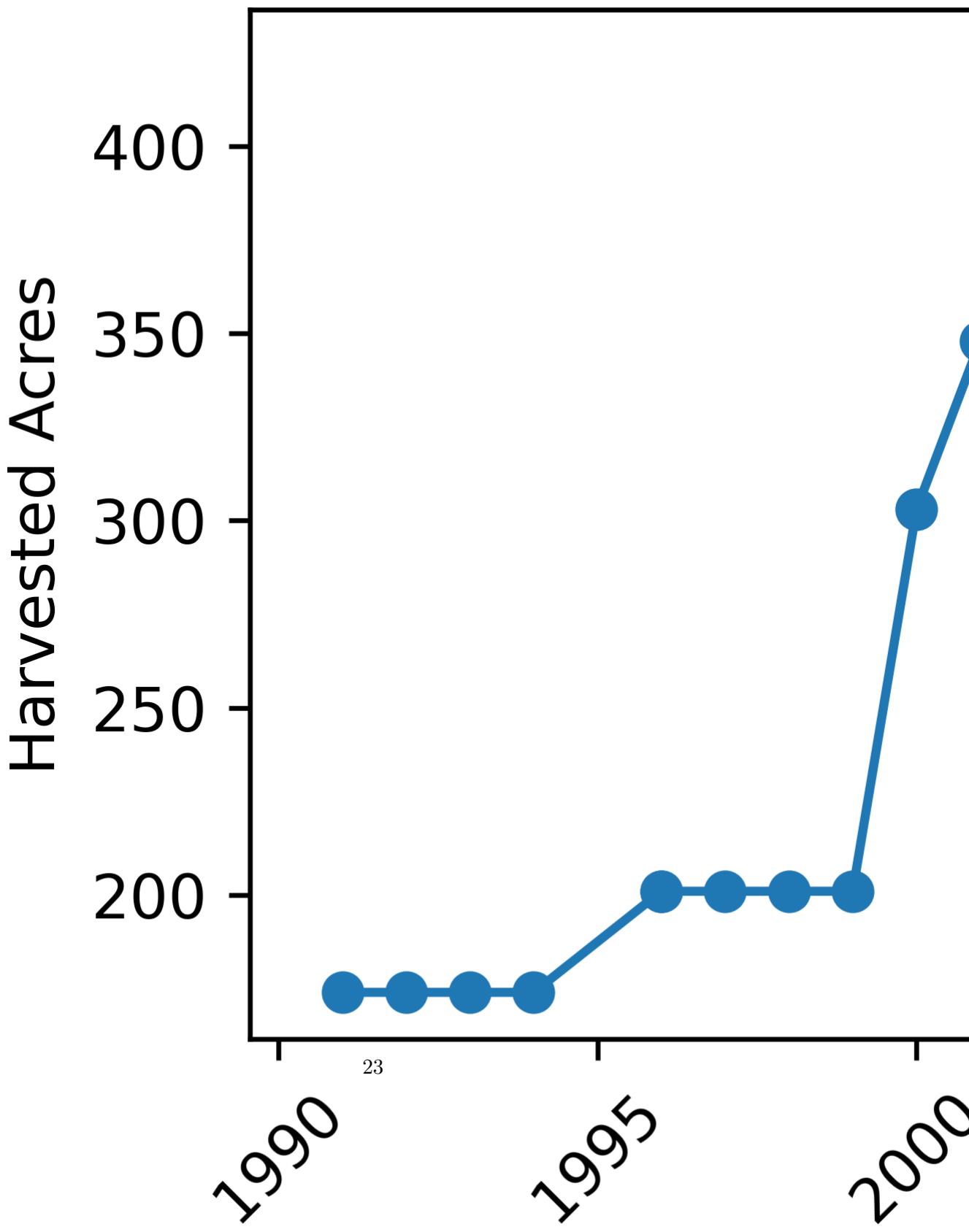
Harvested



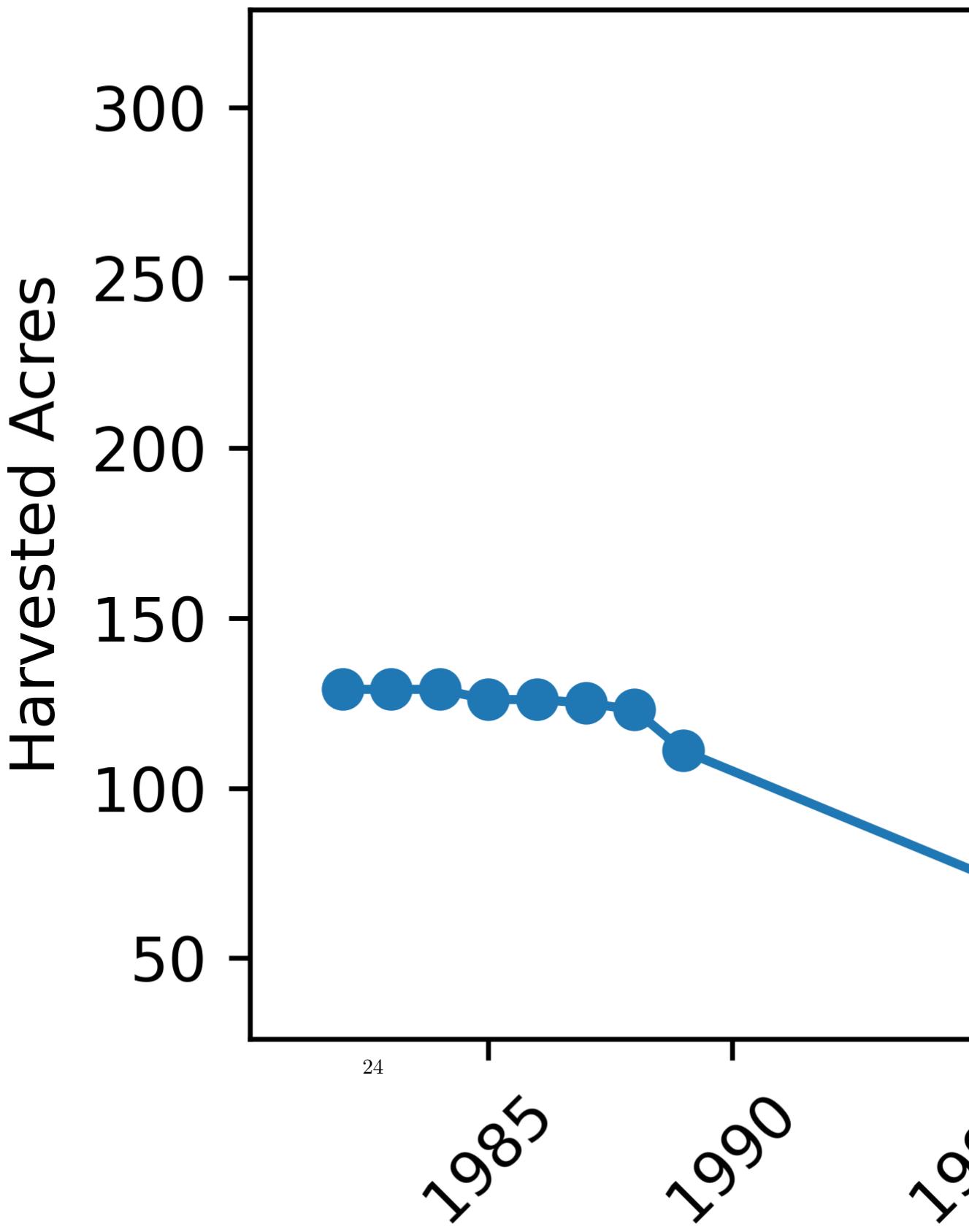
# Harvested Acres



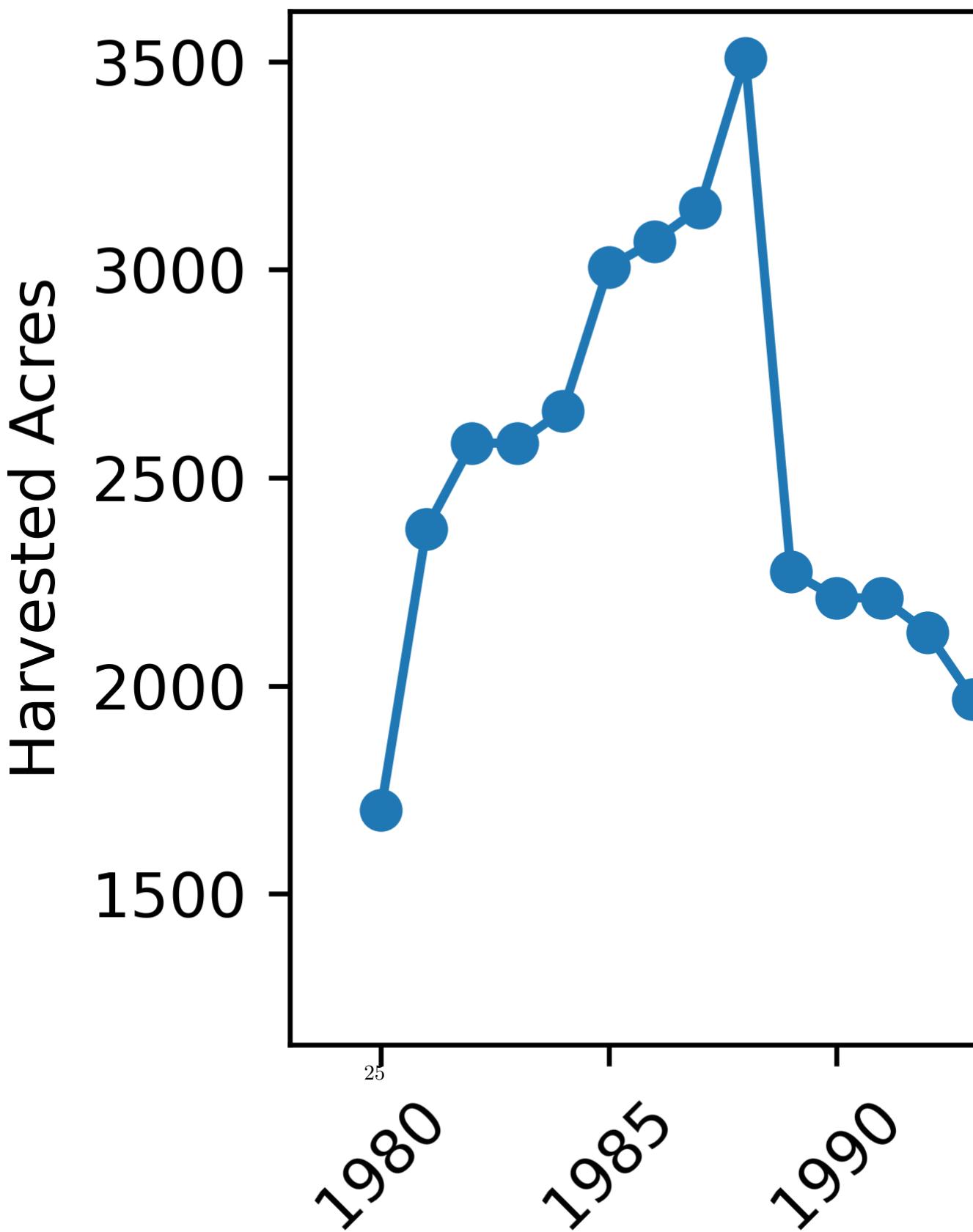
Harvested



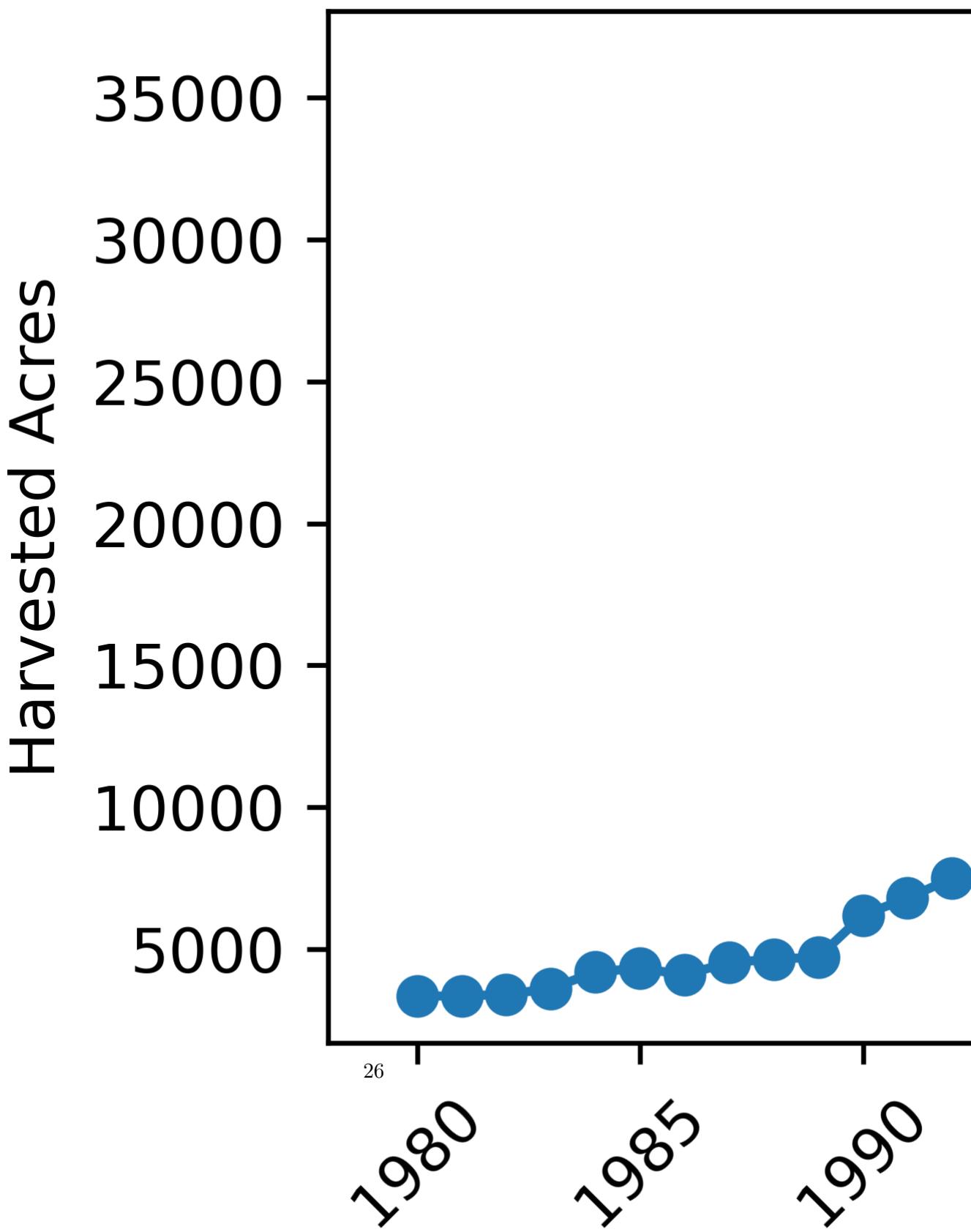
Harvested



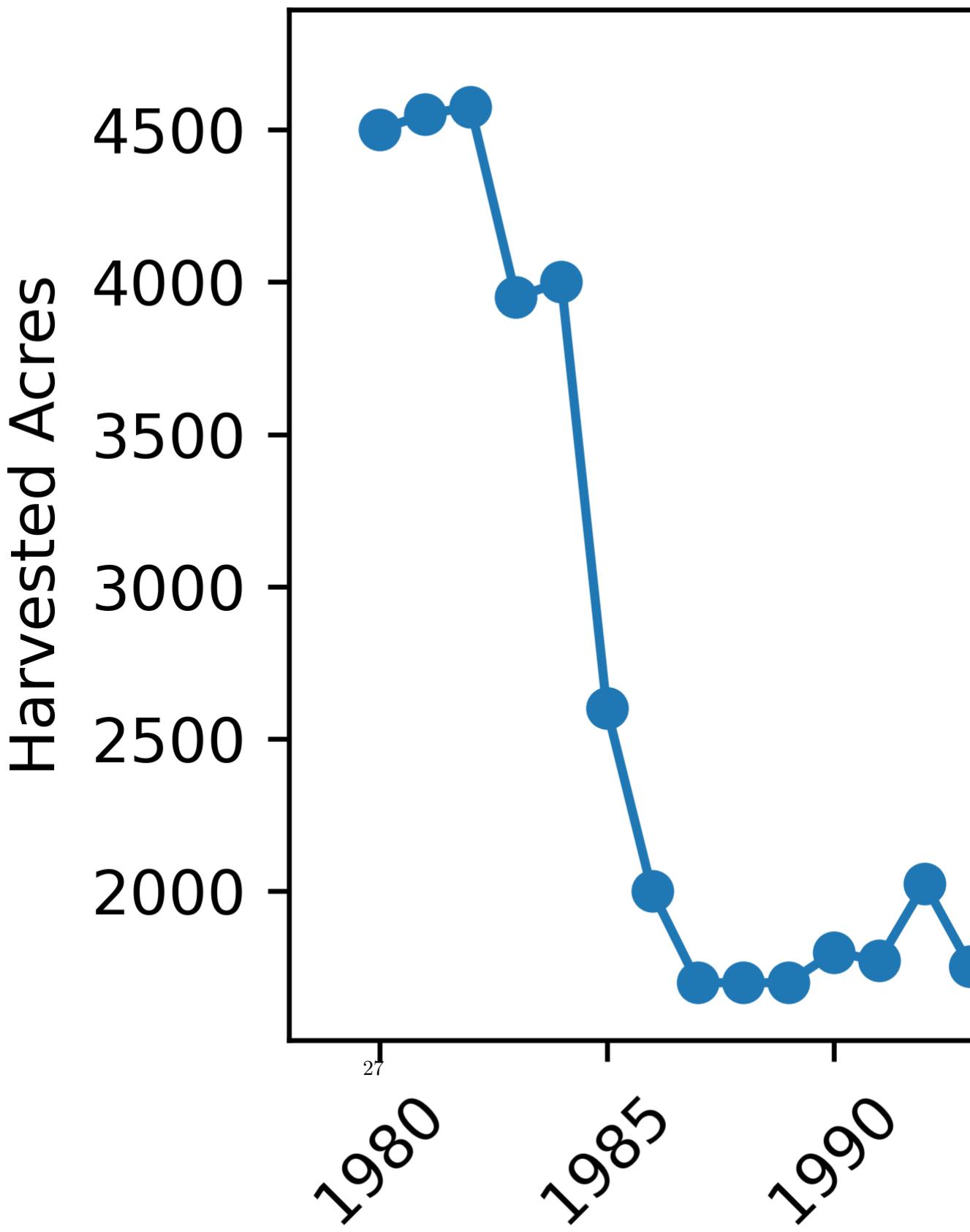
# Harvested Acres



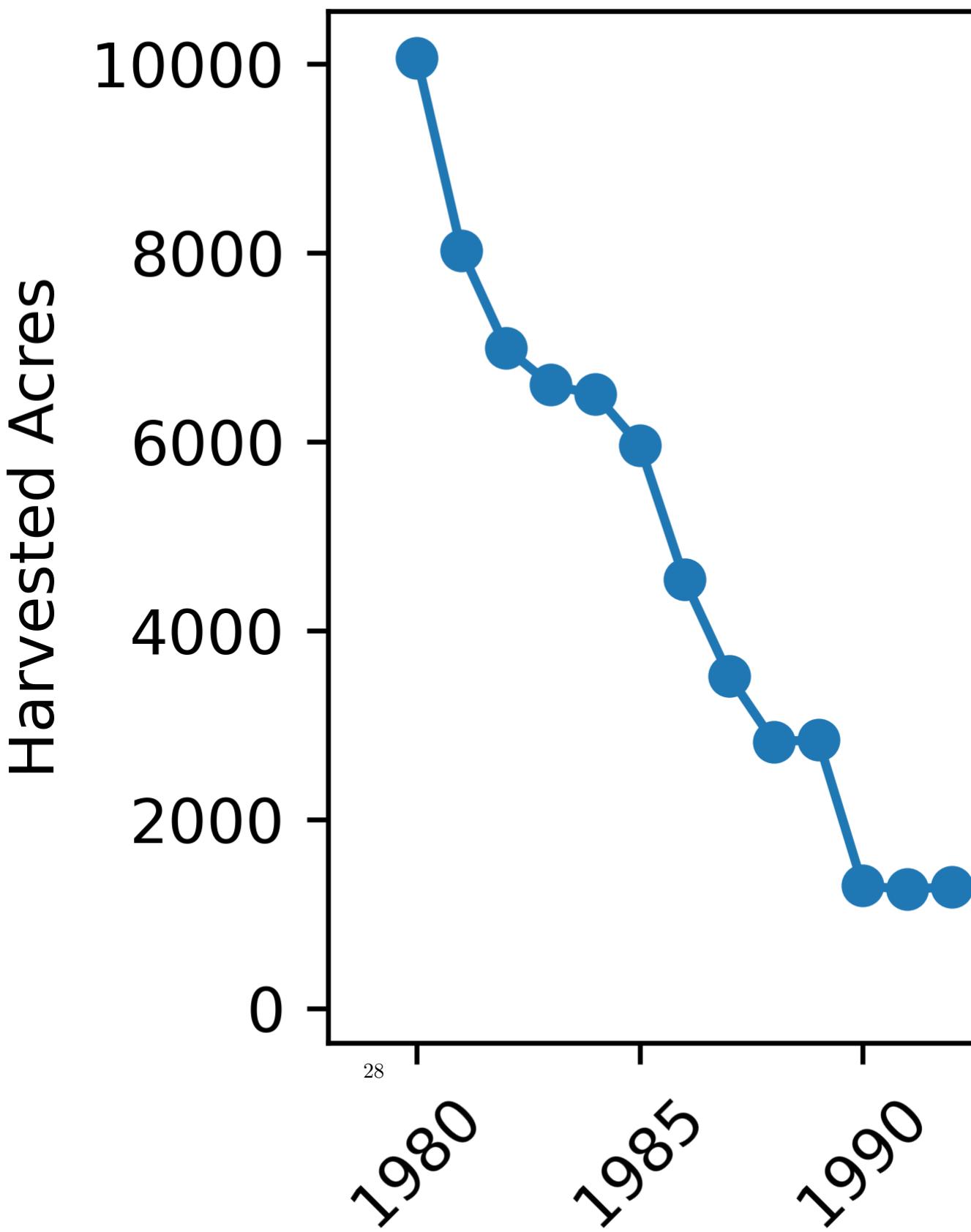
# Harvested A



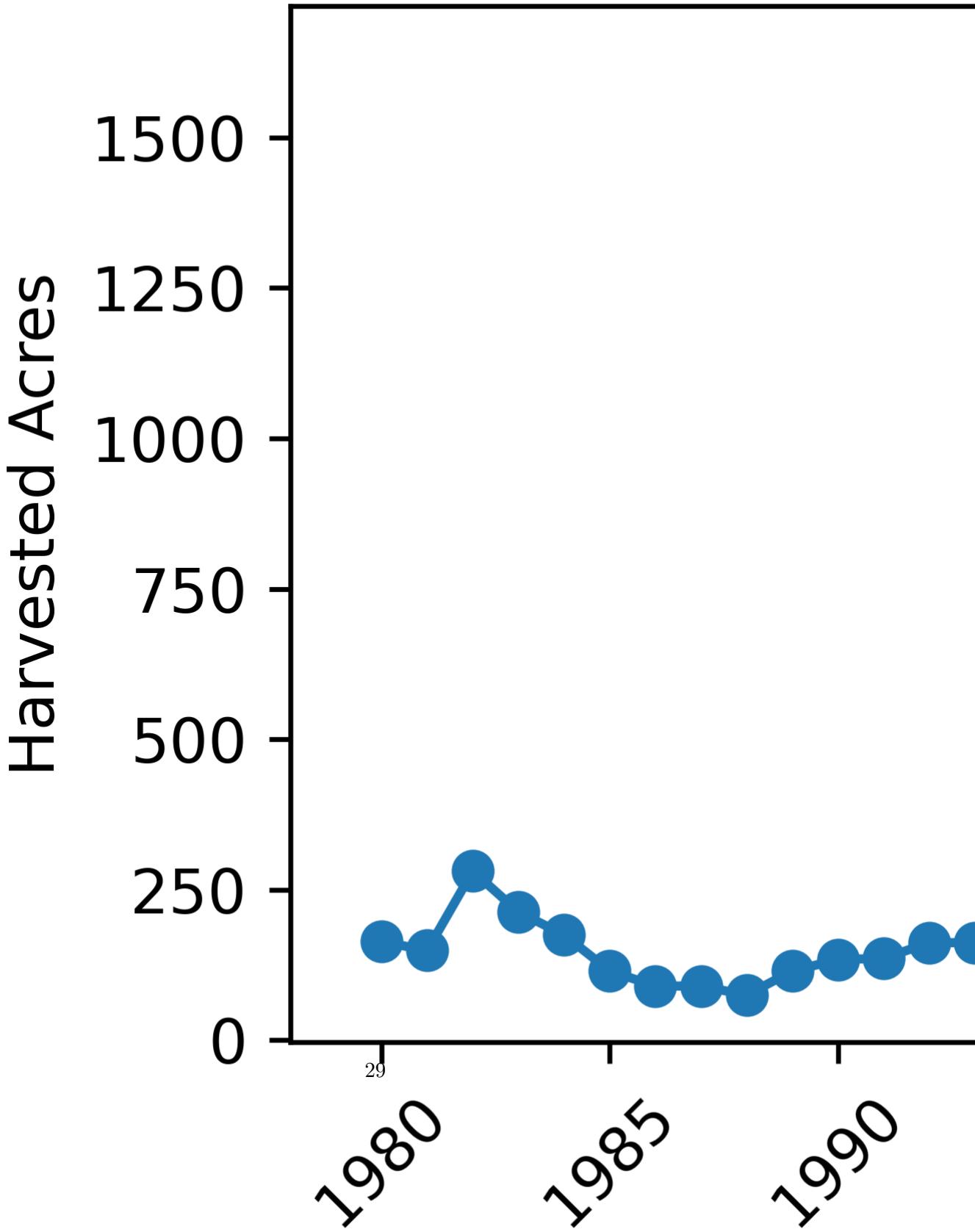
# Harvested Acres



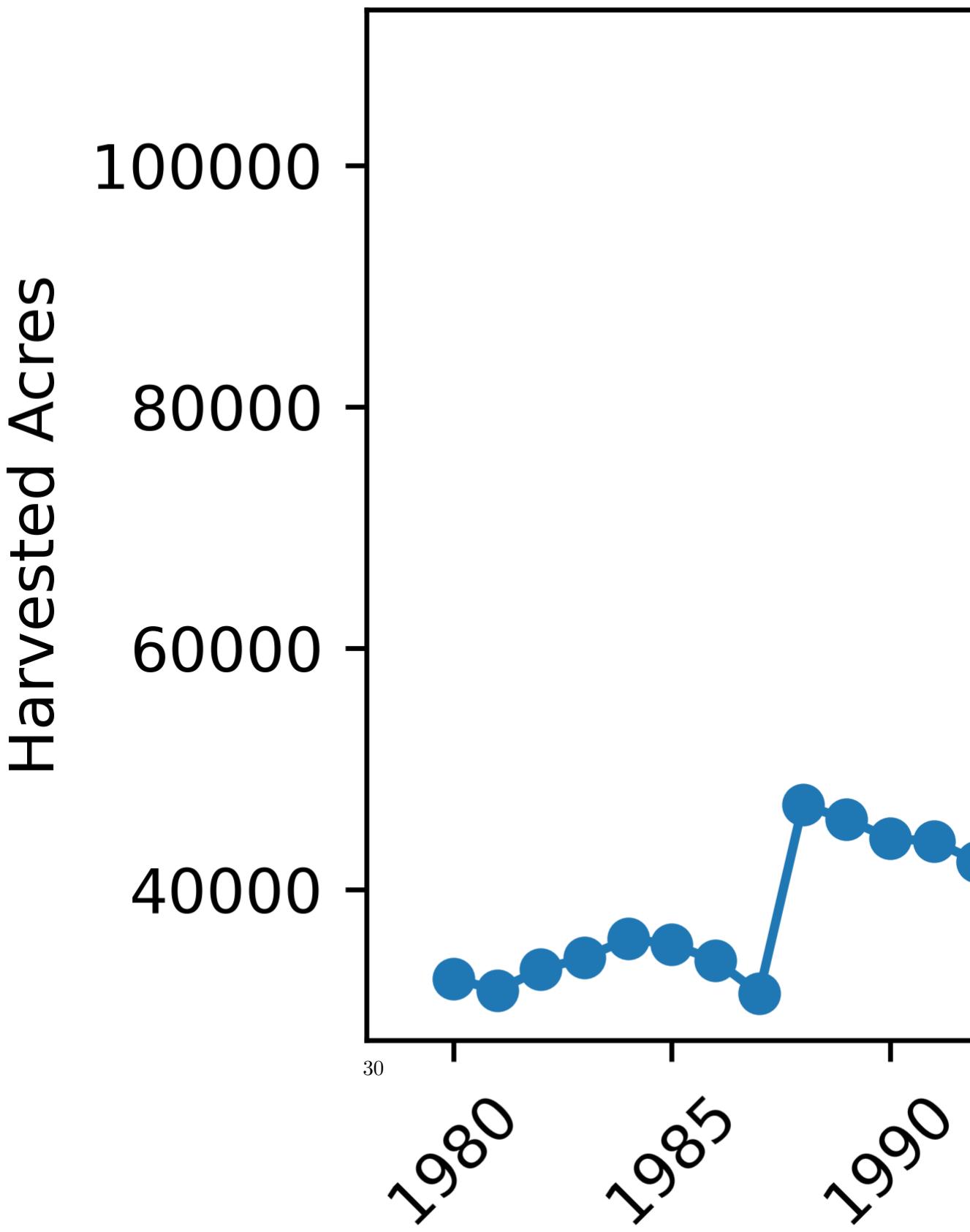
# Harvested Ac



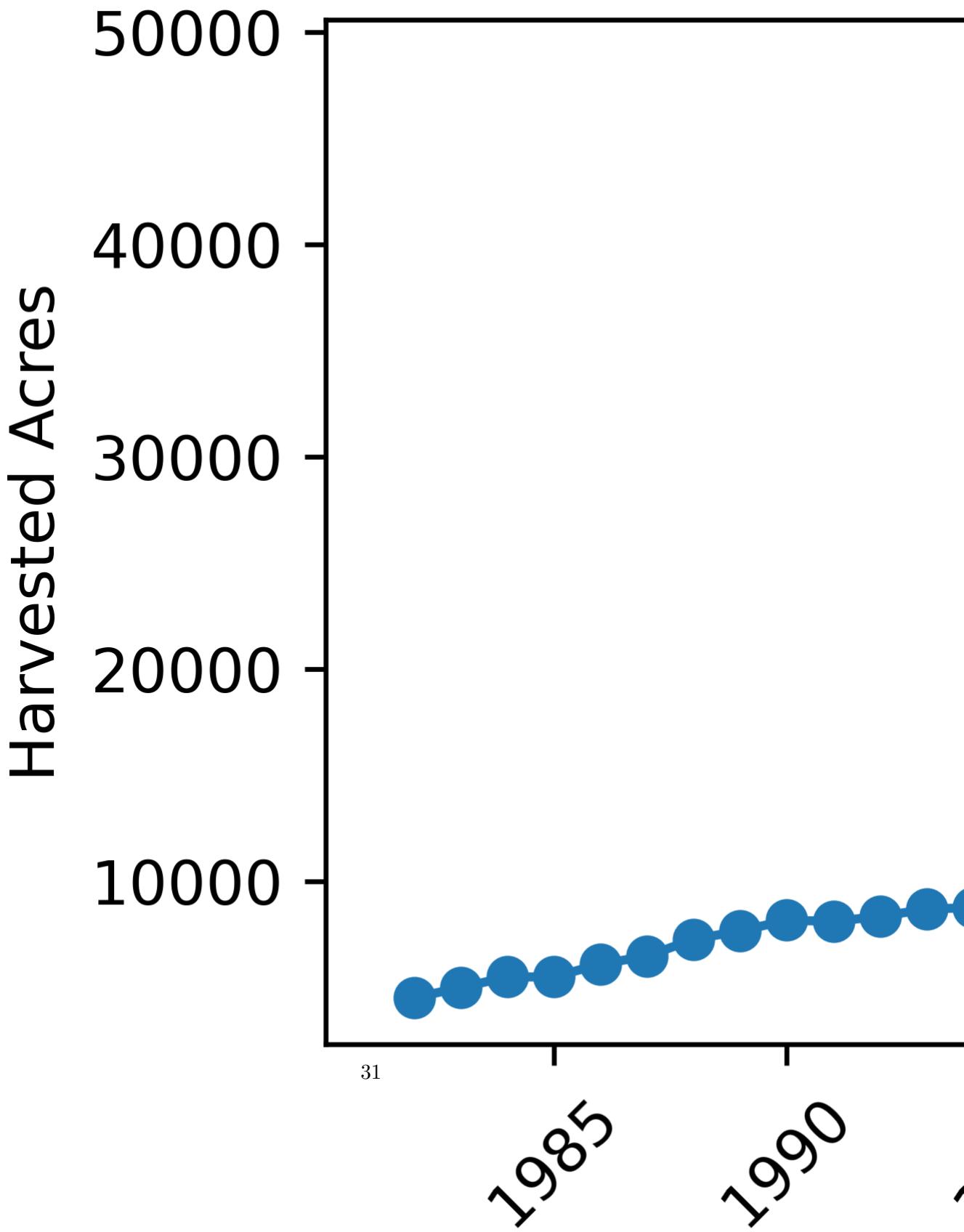
# Harvested Acres



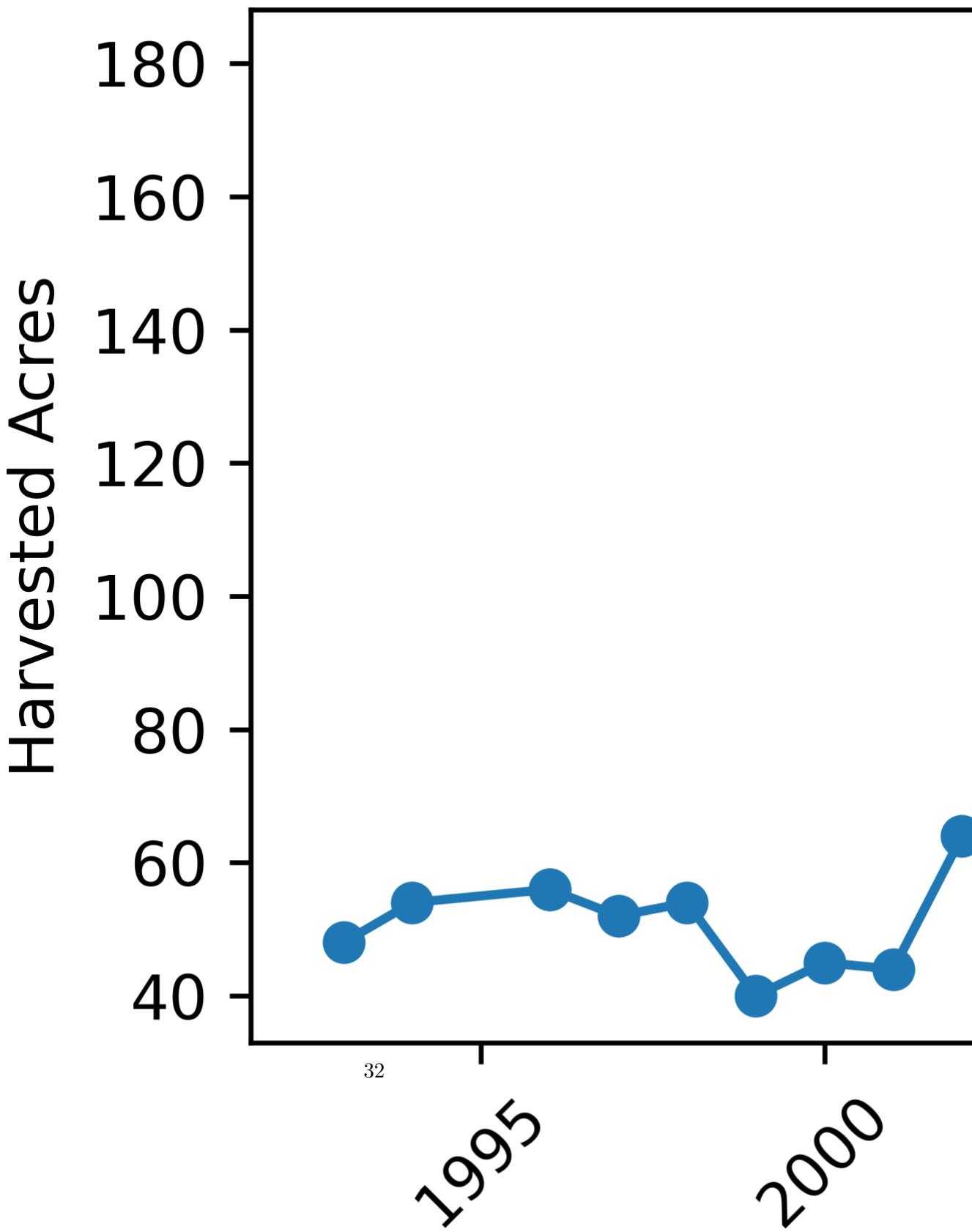
Harvested Acres



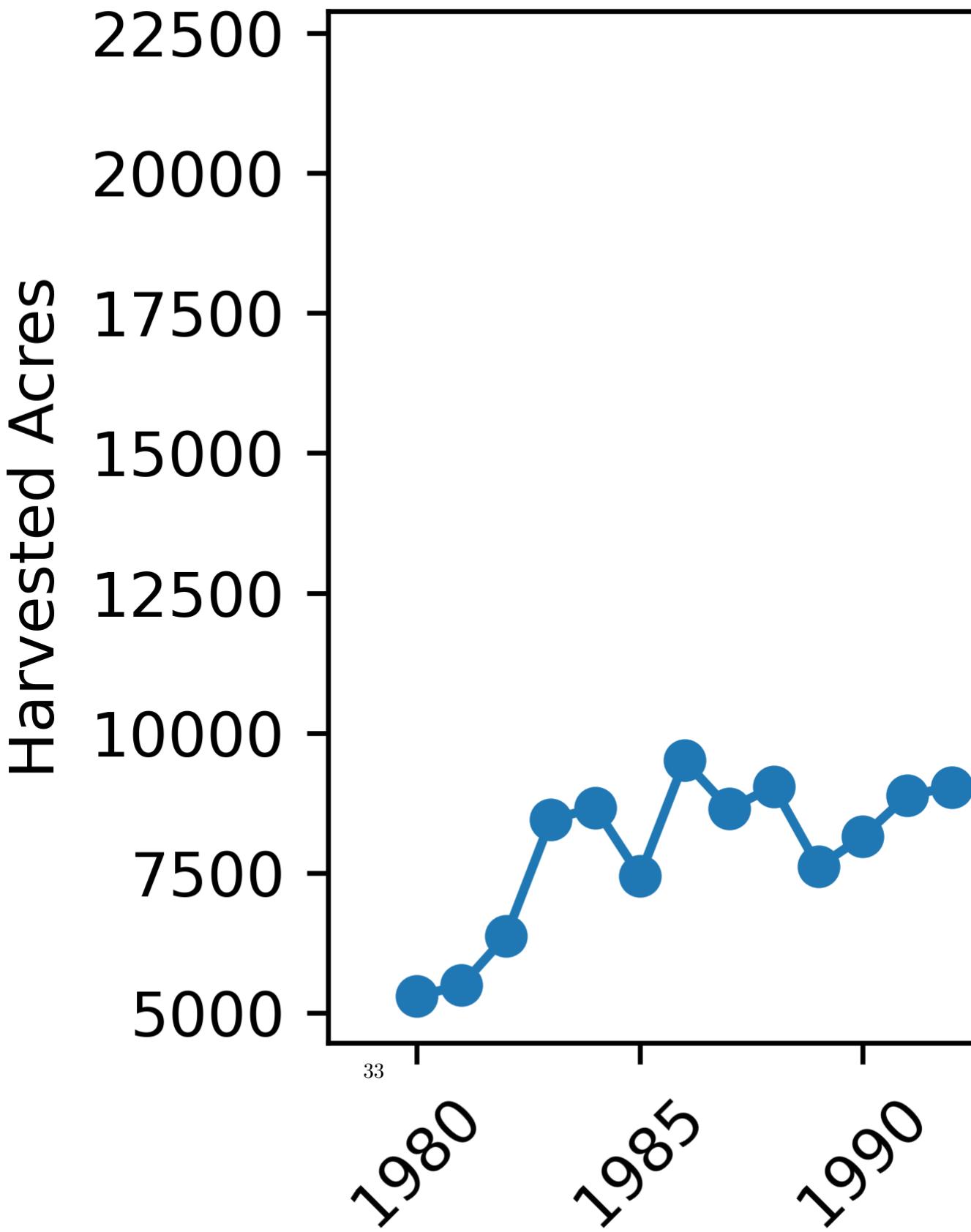
# Harvested Ac



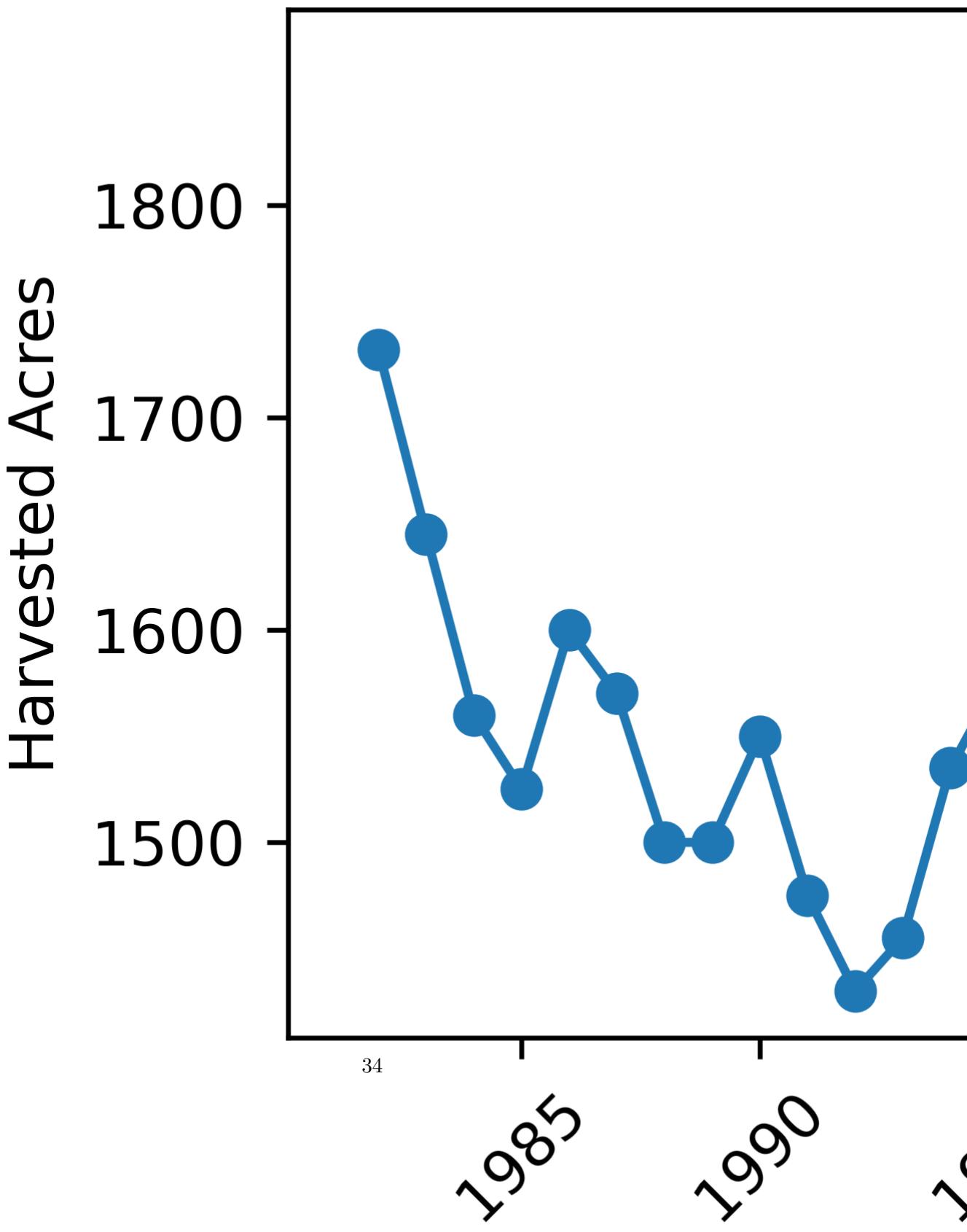
# Harvested A



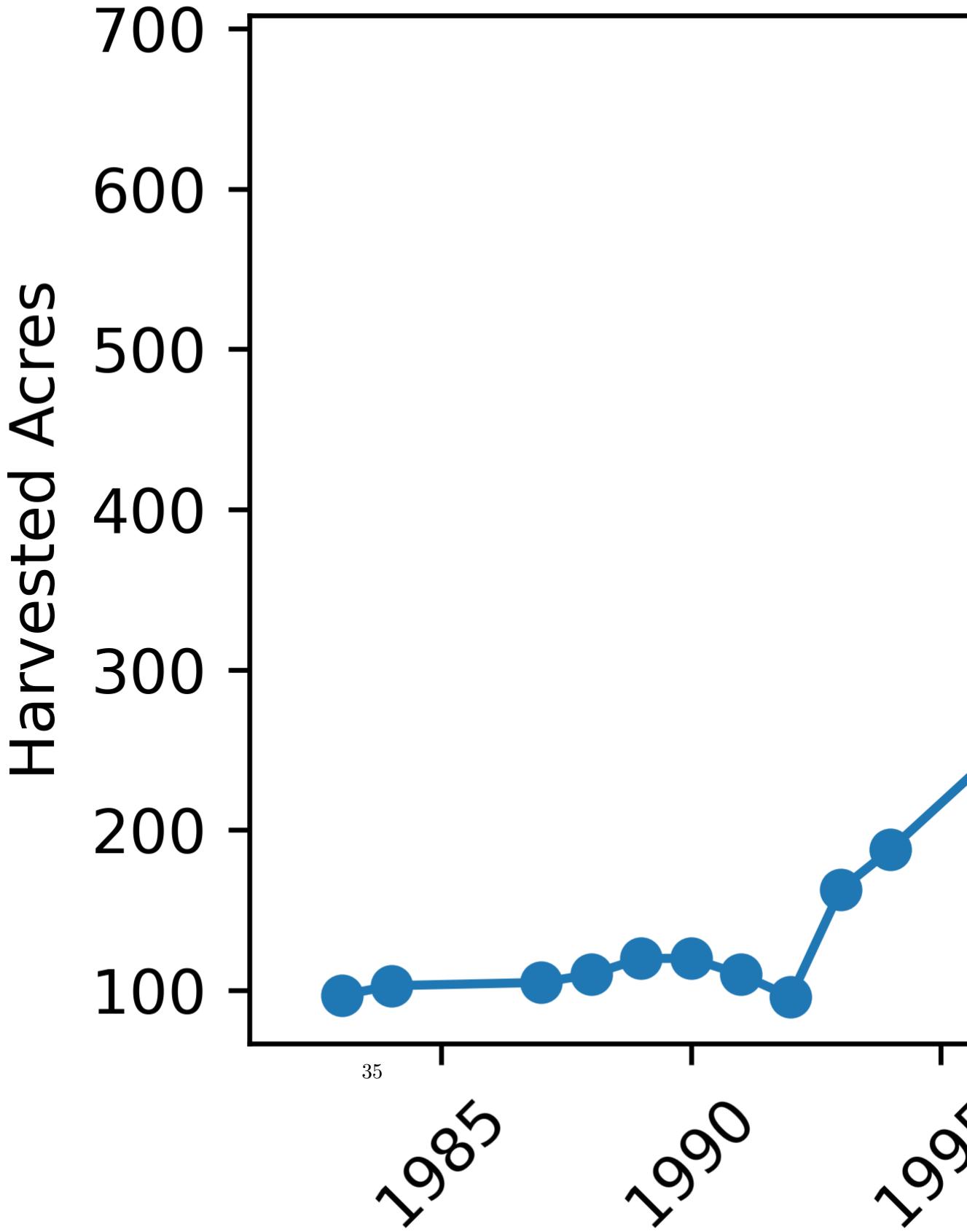
# Harvested Acres



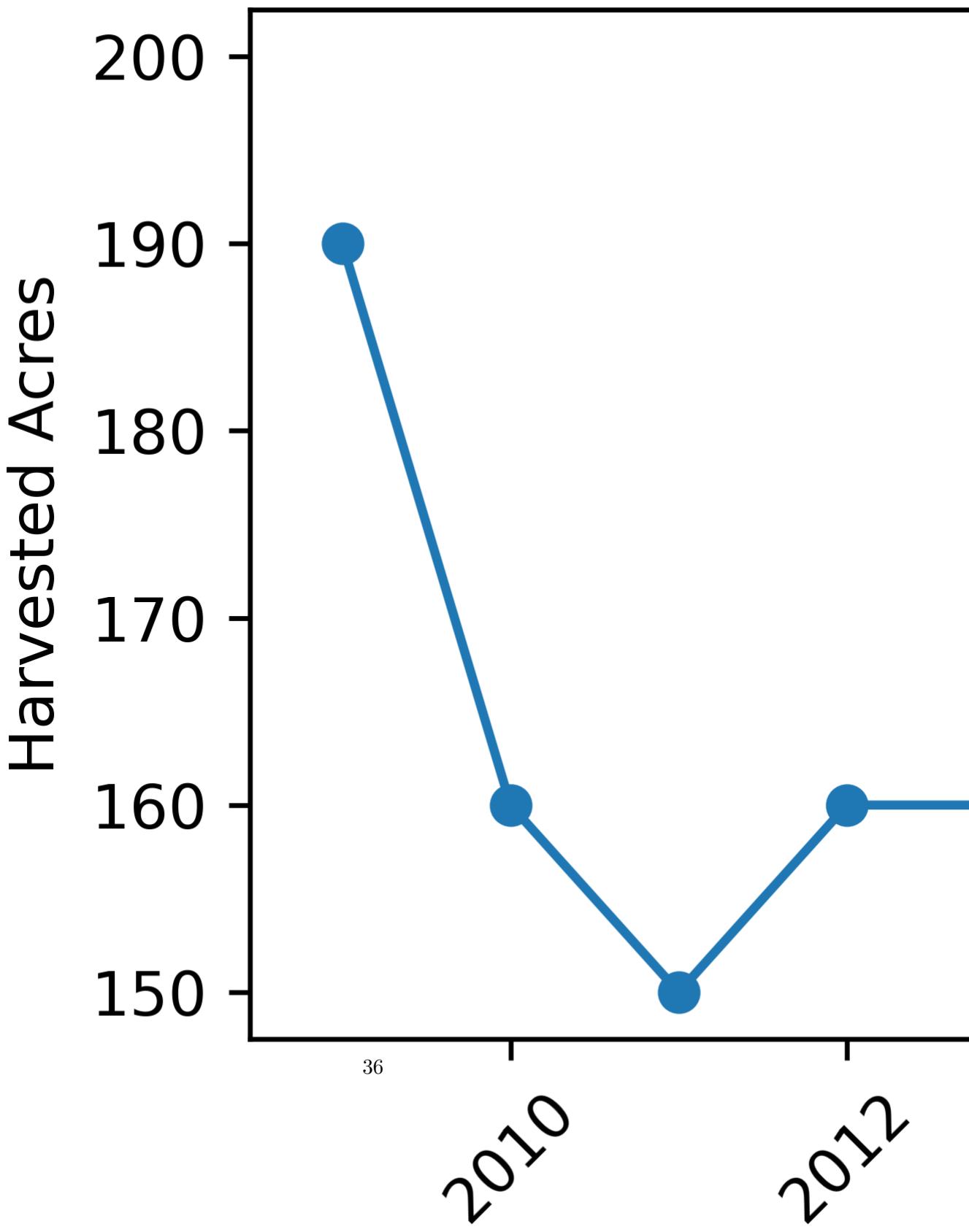
# Harvested Acres



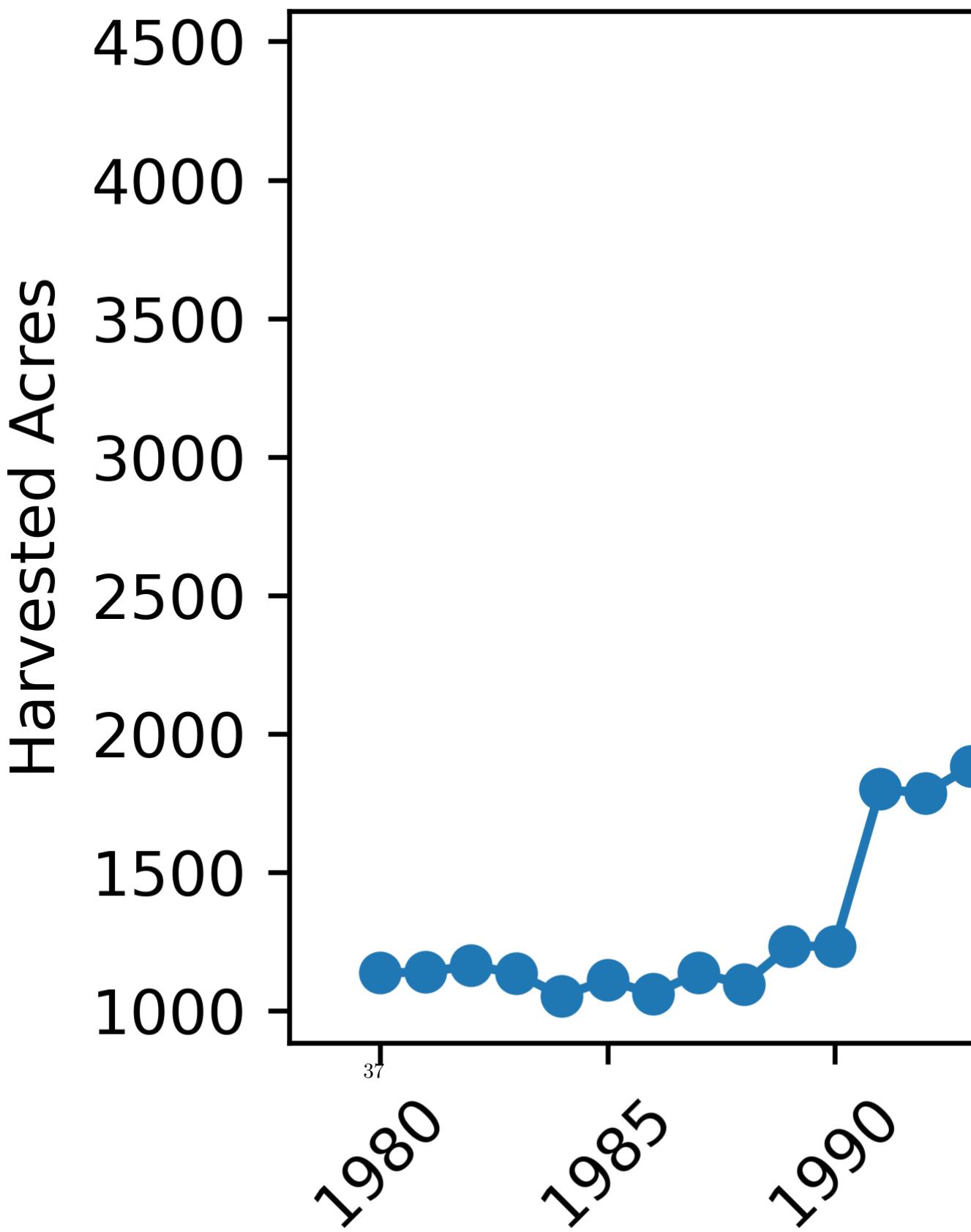
# Harvested A



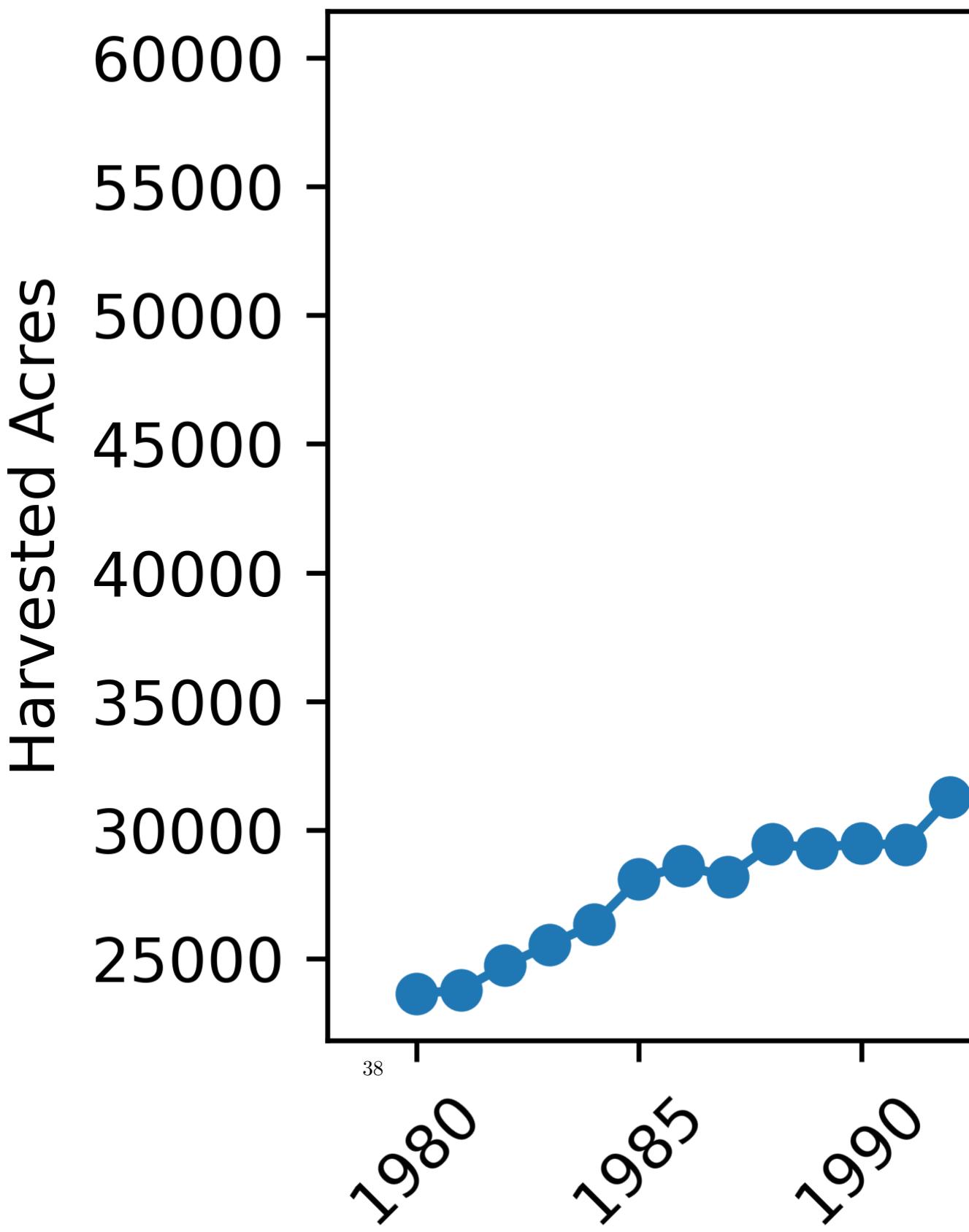
Harvested



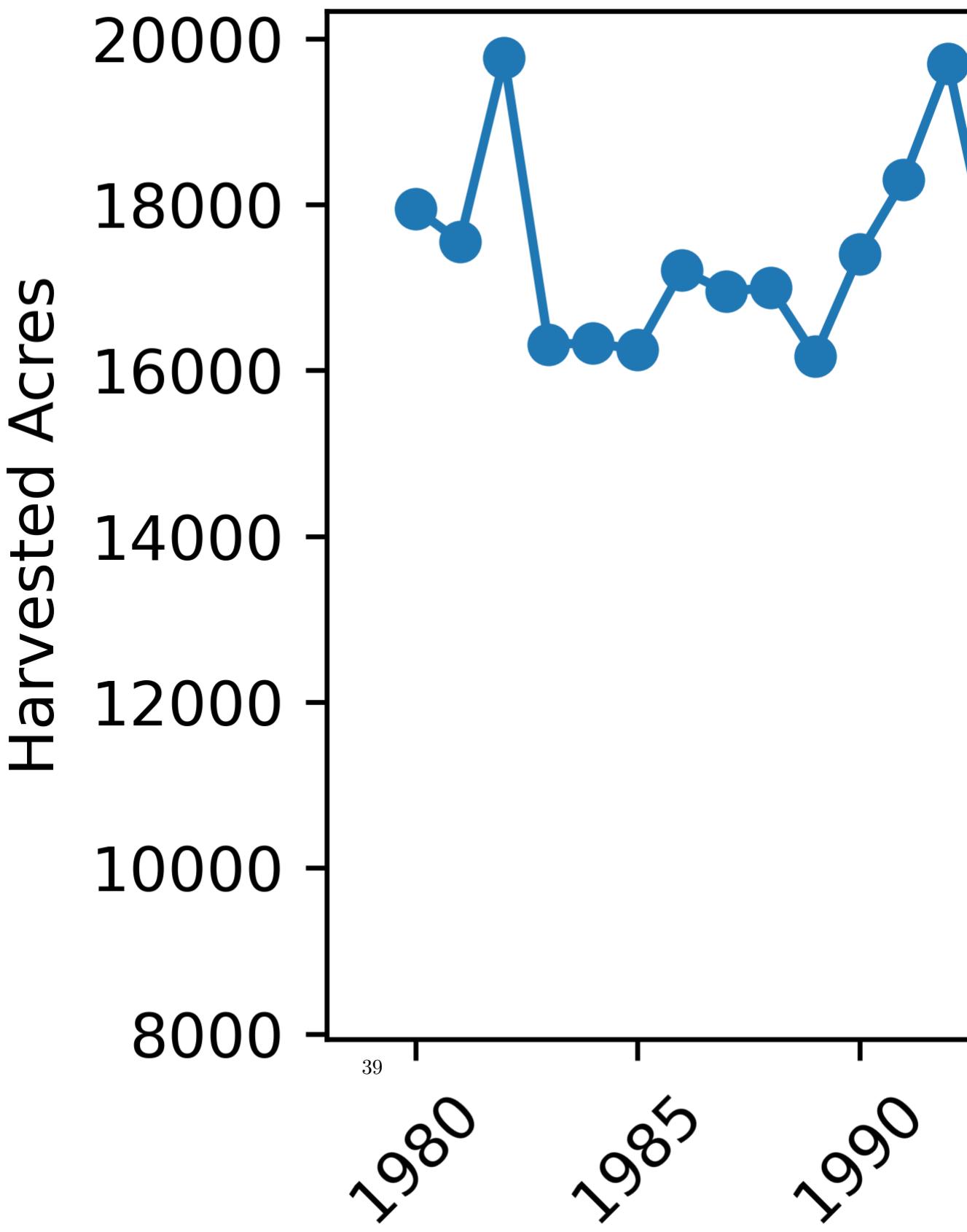
Harvested



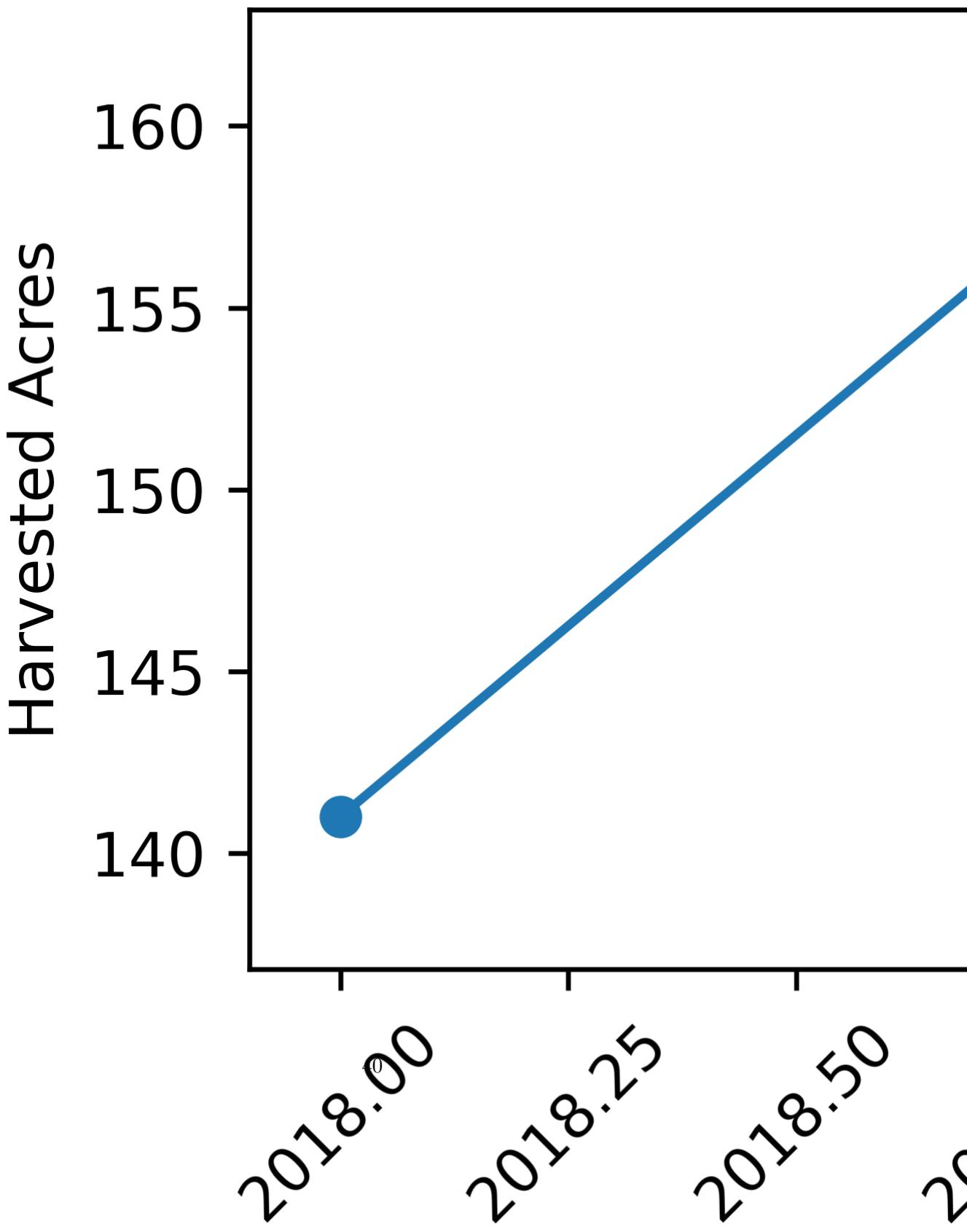
Harvested



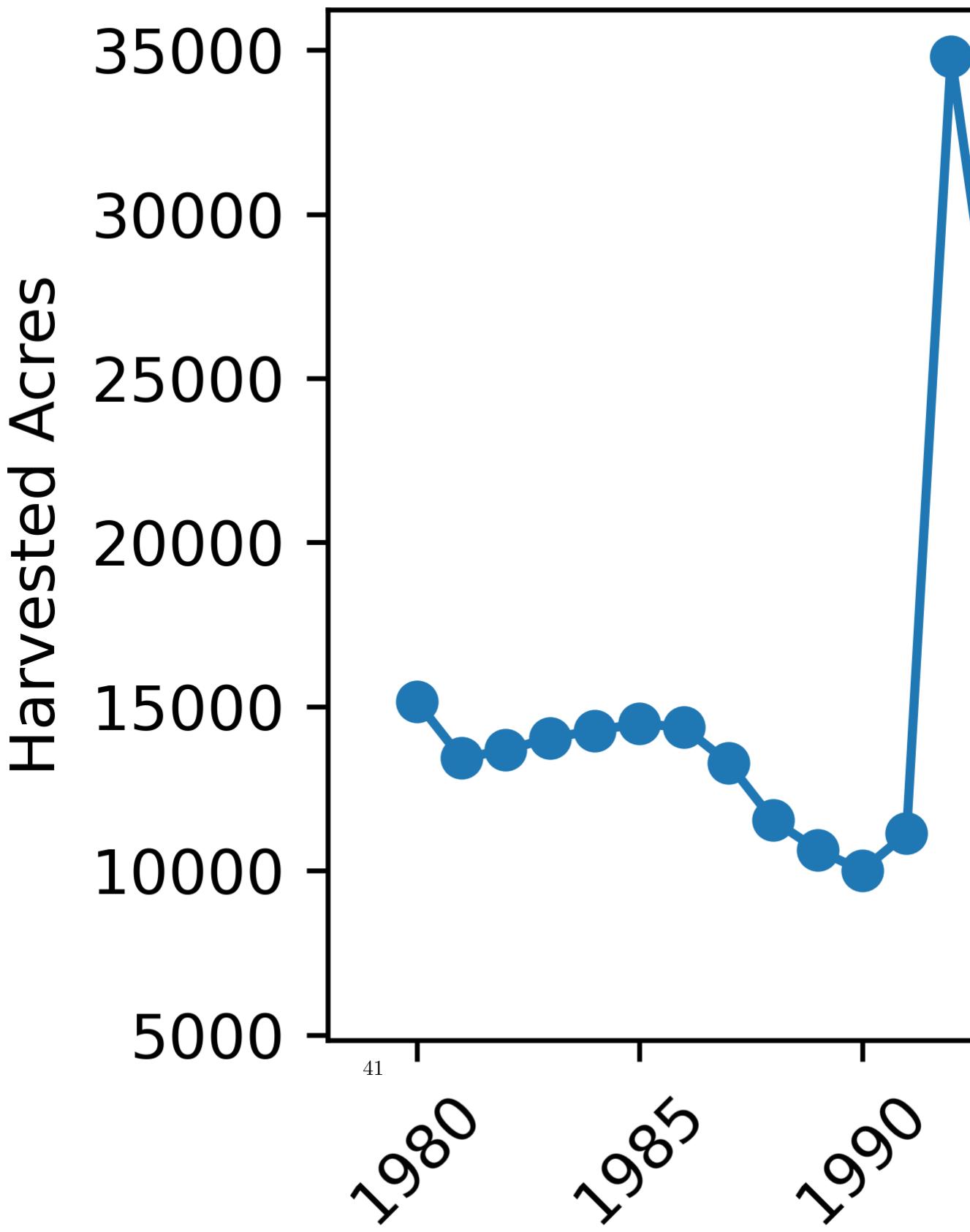
Harvested Acres



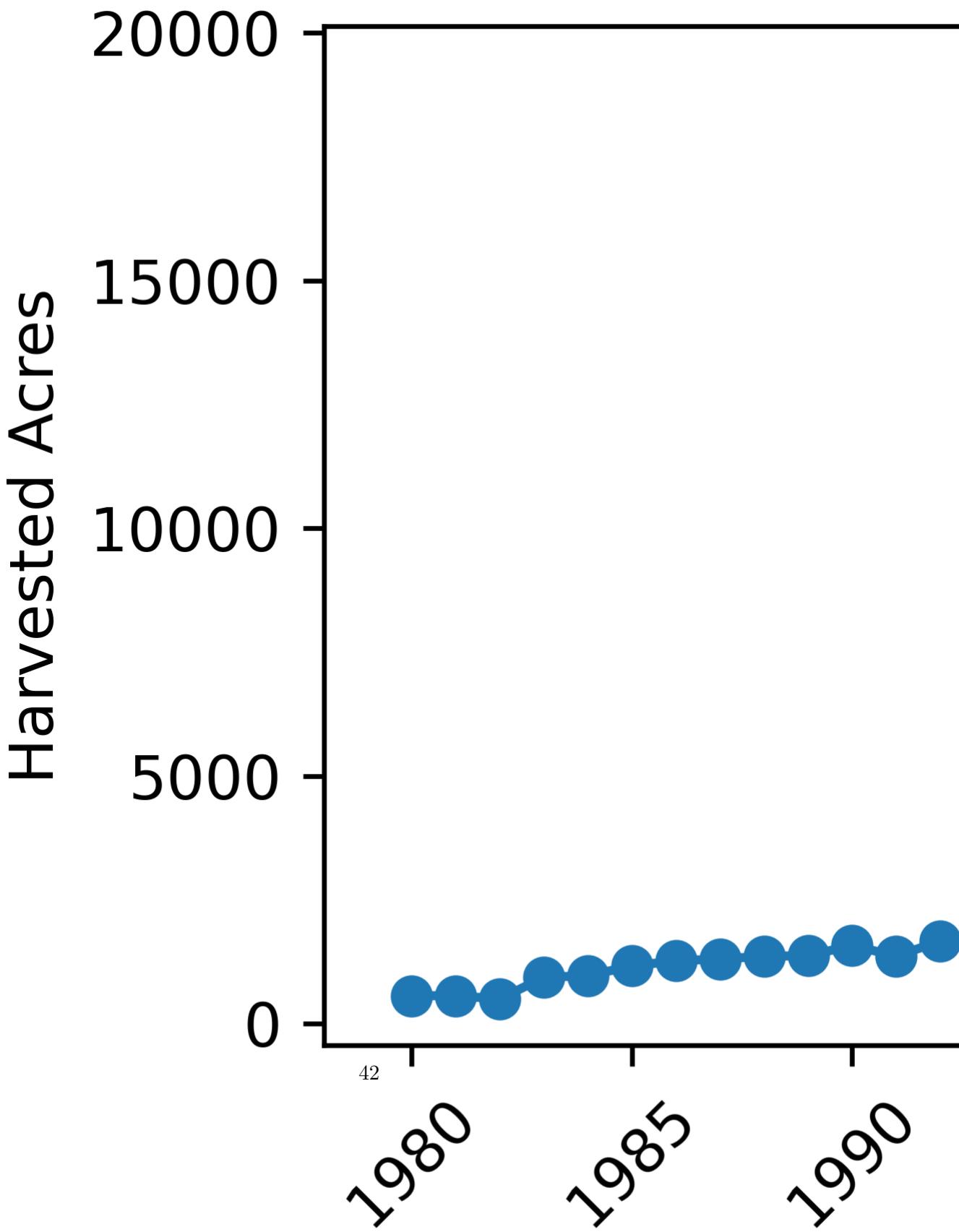
# Harvested Acres



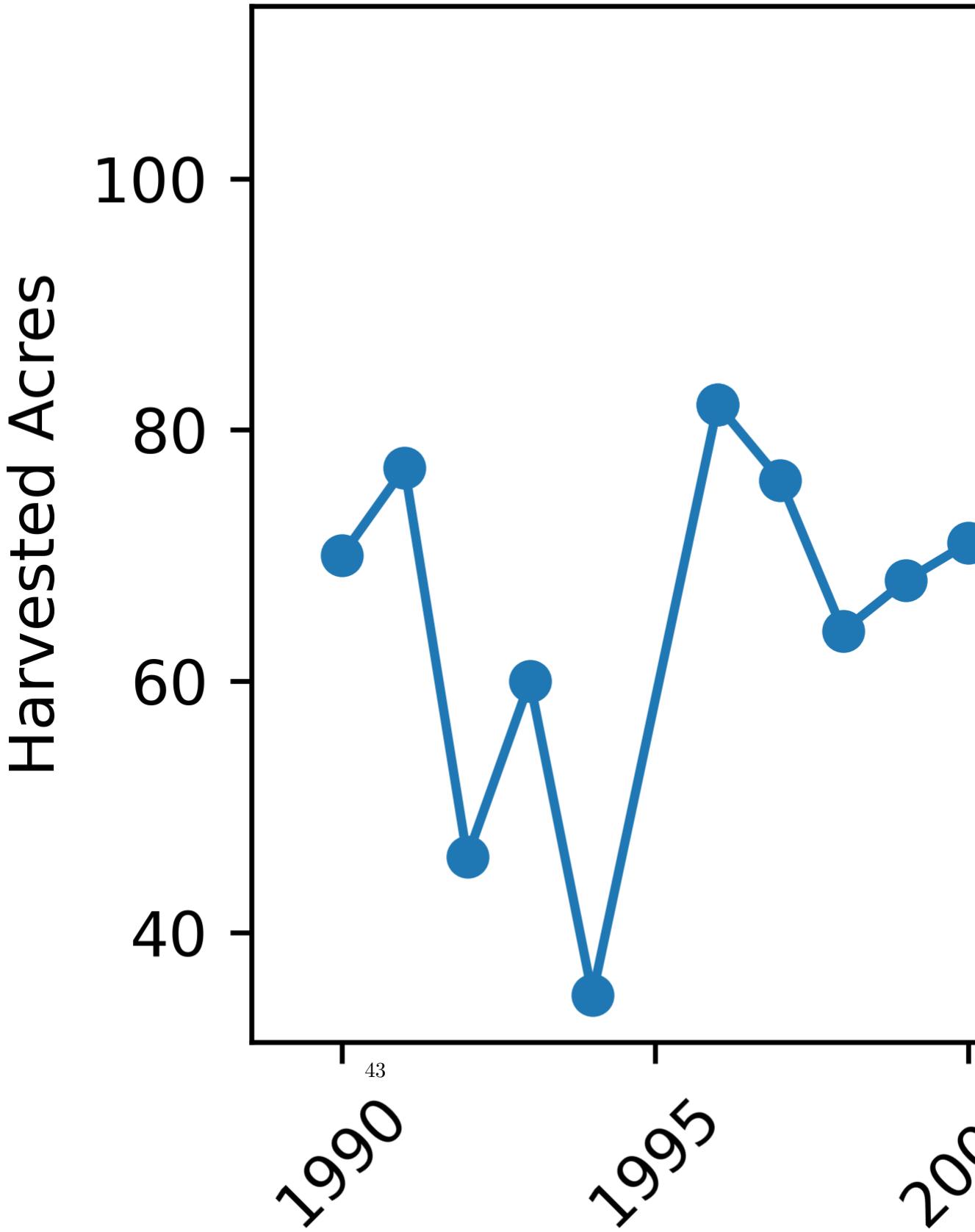
Harveste



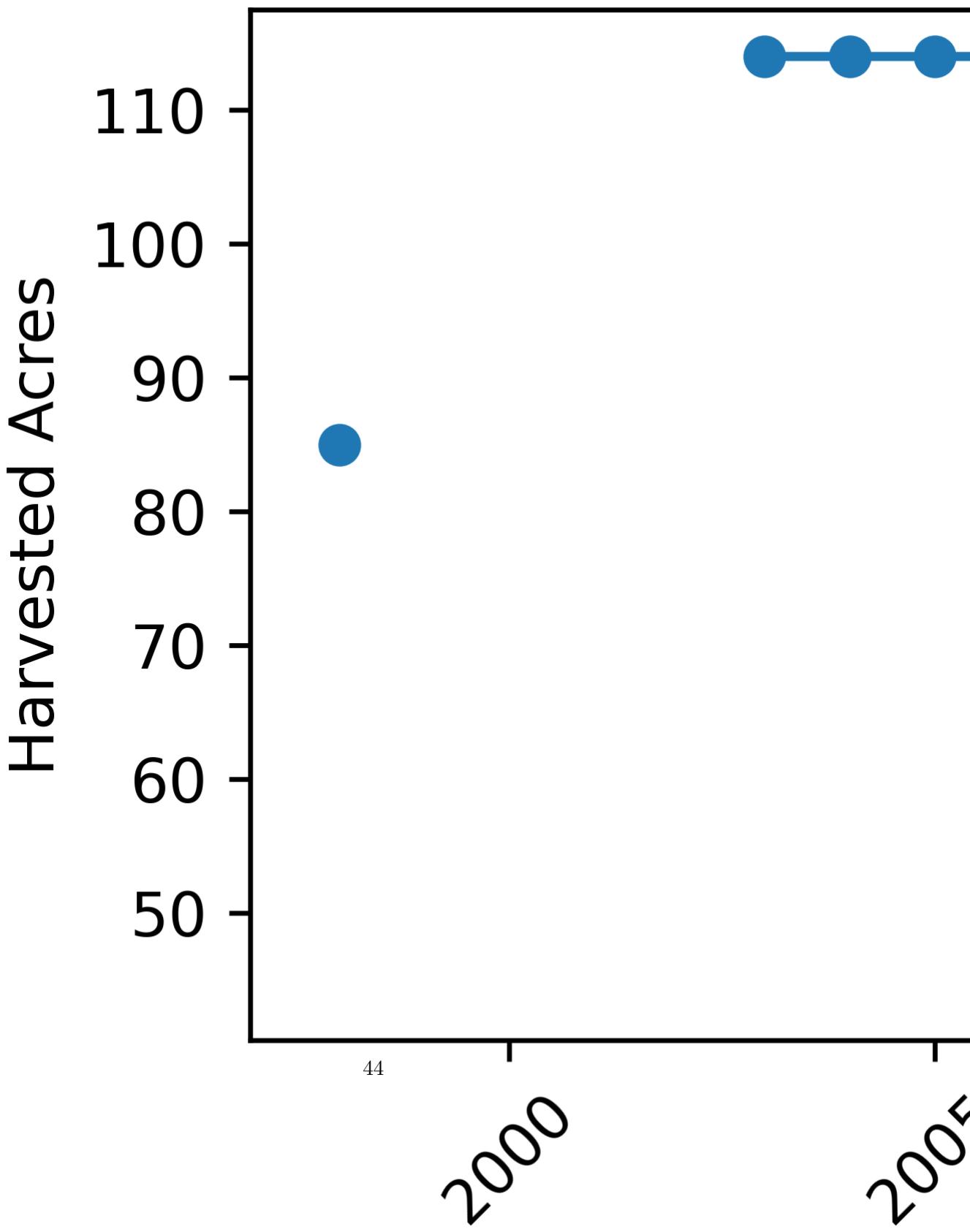
Harvested Acres



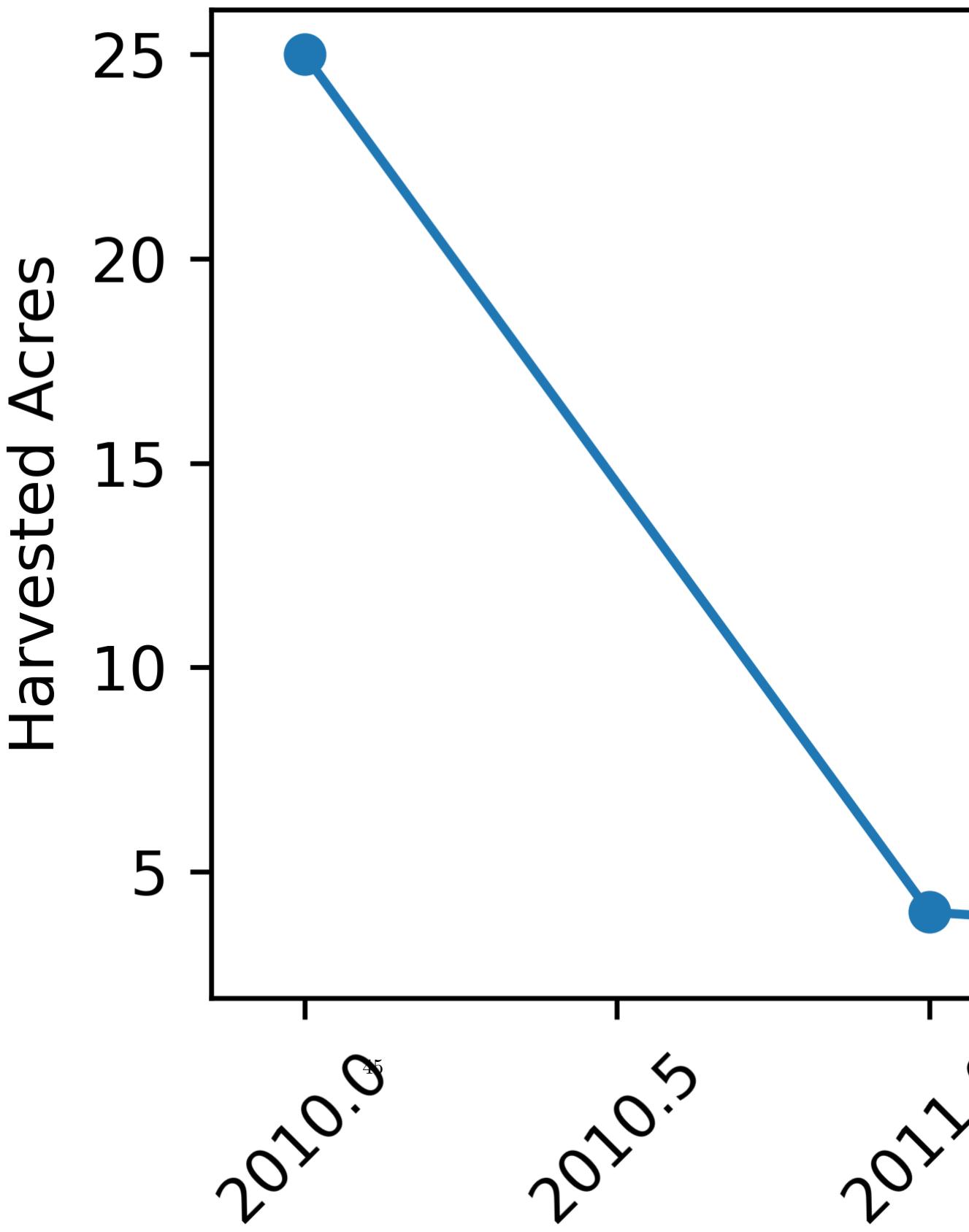
# Harvested Acres



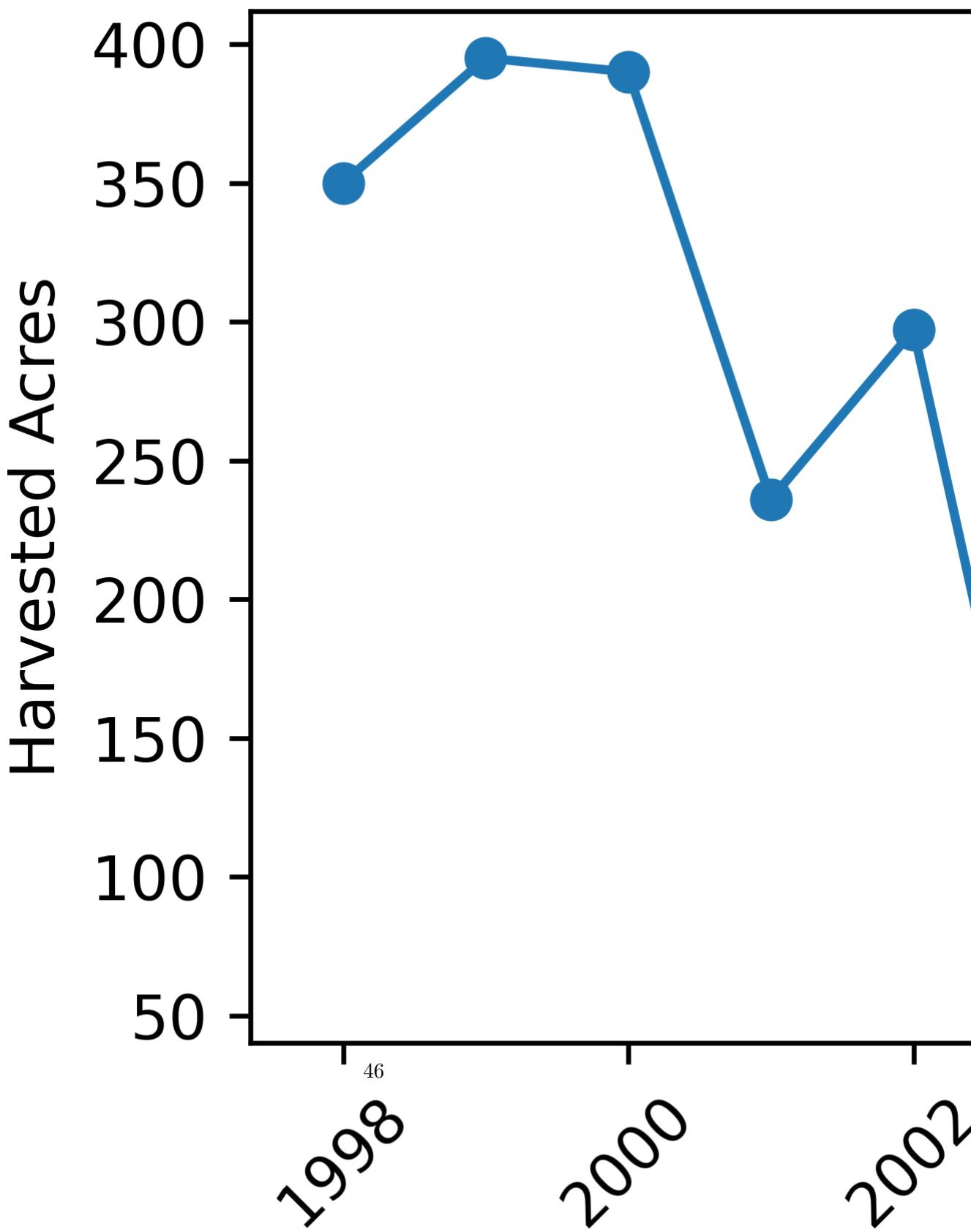
Harvested Acres



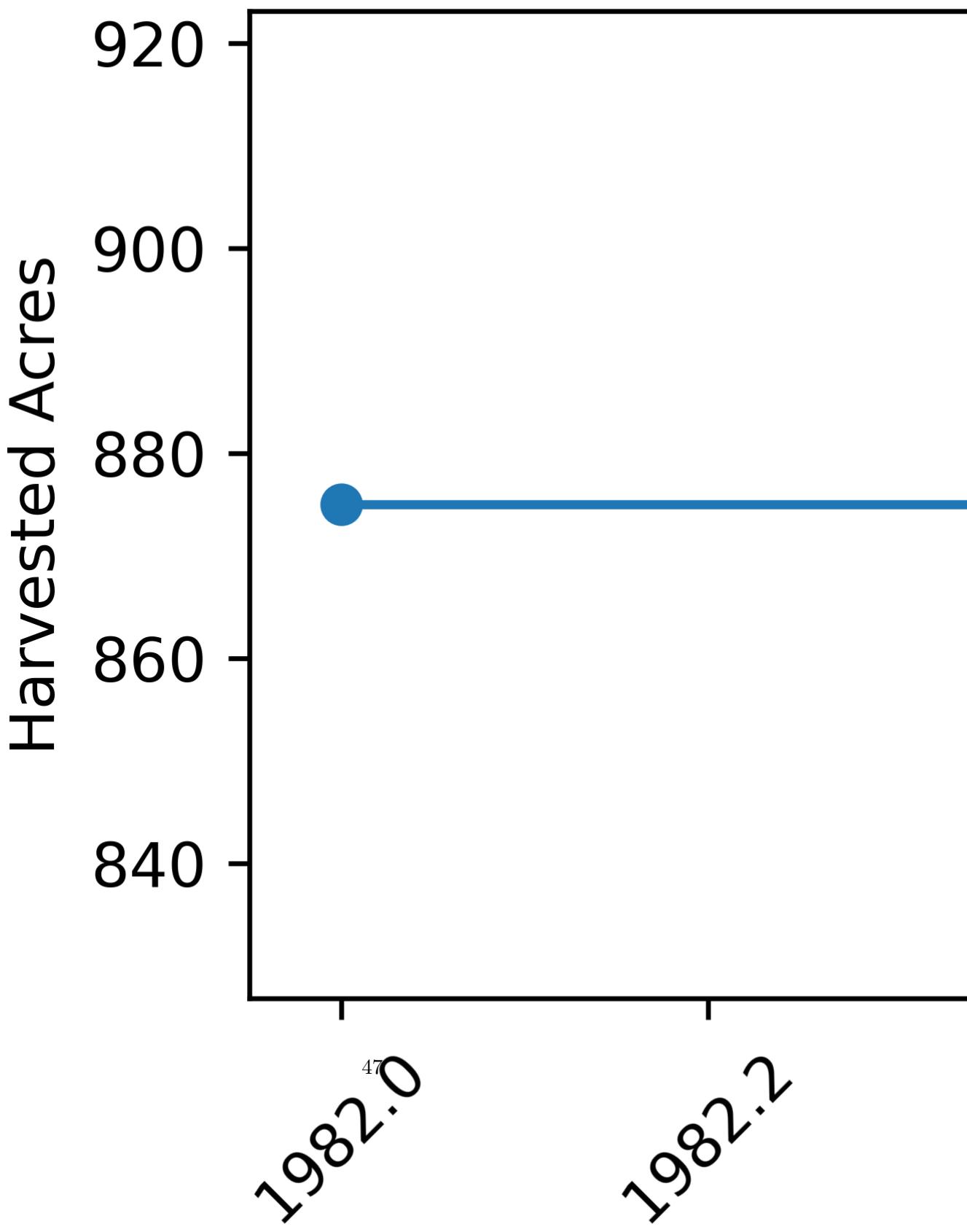
Harvested



Harvested



Harvested



My theory that maybe somehow droughts and heatwaves are effecting yield is based on nothing. There is no patter or corelaion that I can see between the graphs and the dates of major droughts and heatwaves that I can see. At this point i have to a question. Is everything we've heard anecdotal? I mean to say that grape farmers are saying that it's getting harder to keep farming grapes, but so long as they keep watering their grapes, does the heat really impact them at all? I'm gonna ask my farmer friend and see what he tells me