# GALAXY INDEX

*final presentation*

*1.29.13*

# PROBLEM

existing consumer financial indexes are poll-based and monthly. this precludes up-to-the-minute analysis and prediction of market behaviour and trends.

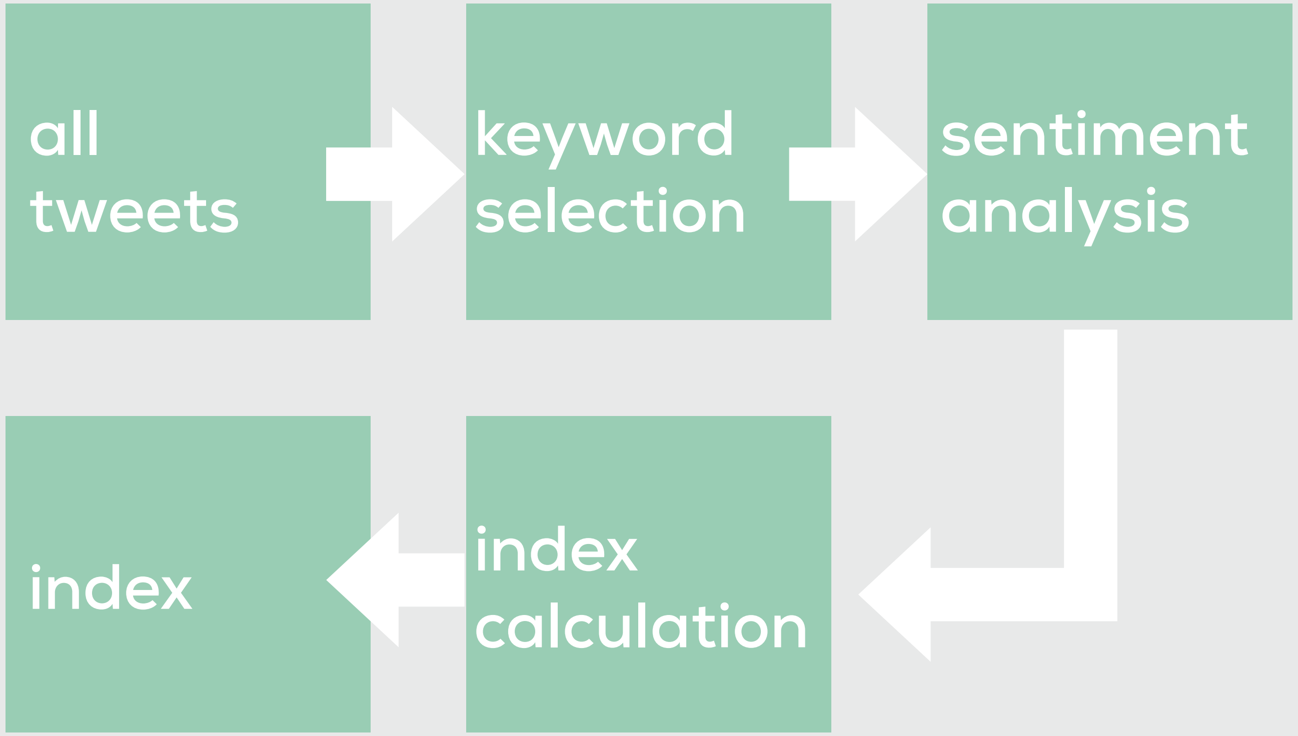
# GOAL

develop a supplemental financial index describing consumer sentiment by mining social media data

# TEAM

rohan kulkarni -mit
sascha boheme- mit
philipp staiger- mit
arindra das- helsinki
udit anand- helsinki
andrei simonchyk- cologne
nathan sundberg- scad

# PROCESS



all tweets

keyword selection

sentiment analysis

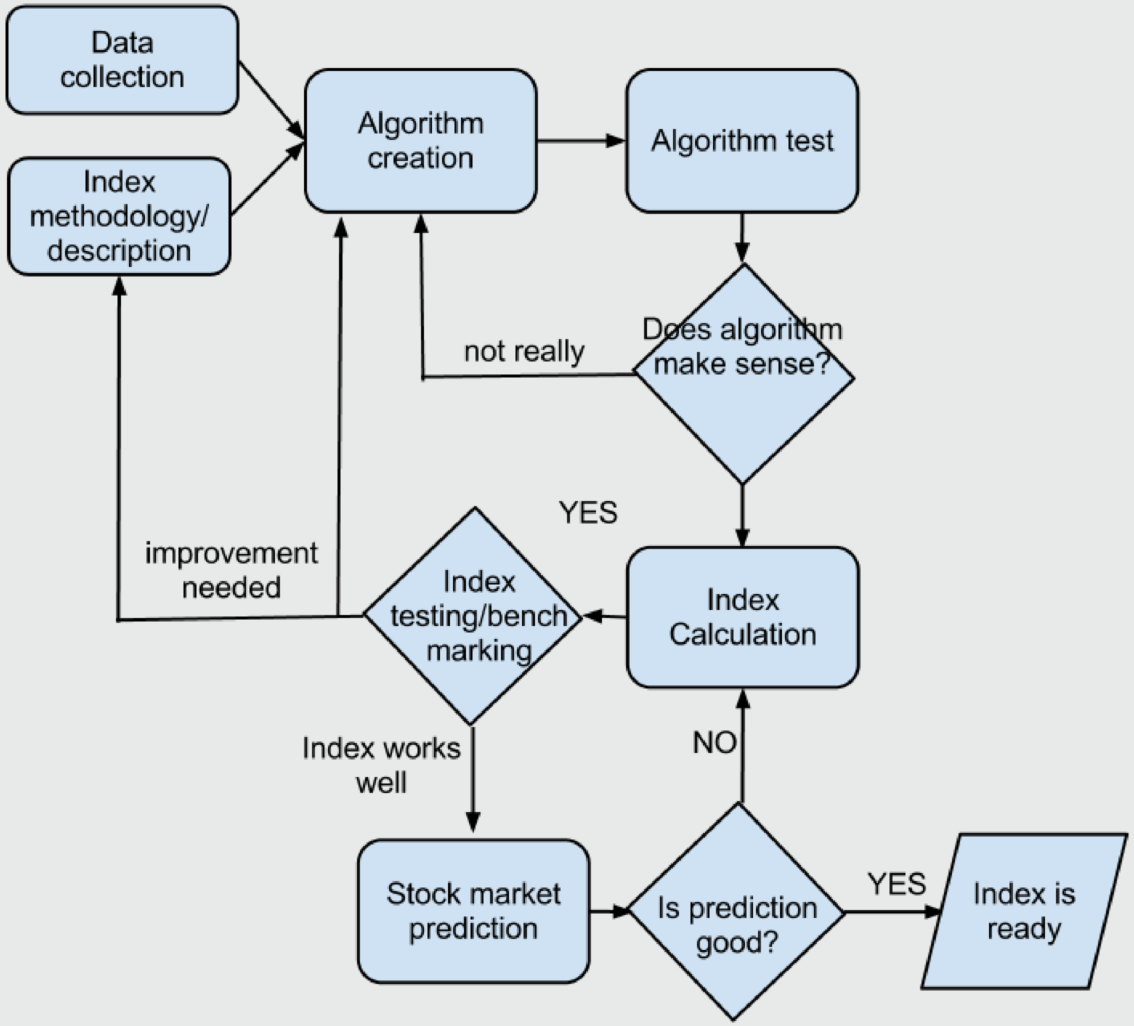
index calculation

index

# ELEMENTS

1) gather twitter data
2) index methodology/concept
3) algorithm creation
4) index calculation and benchmarking
5) financial market prediction
6) communication methdology

# FLOW CHART


Data collection
Index methodology/ description
Algorithm creation
Algorithm test
Does algorithm make sense?
not really
YES
Index Calculation
Index testing/bench marking
improvement needed
Index works well
Stock market prediction
Is prediction good?
NO
YES
Index is ready

# RESEARCH

twitter data collection

4 GB of Twitter data from AWS Condor search running every 20 mins

4 million tweets from Kang Zhang

# RESEARCH

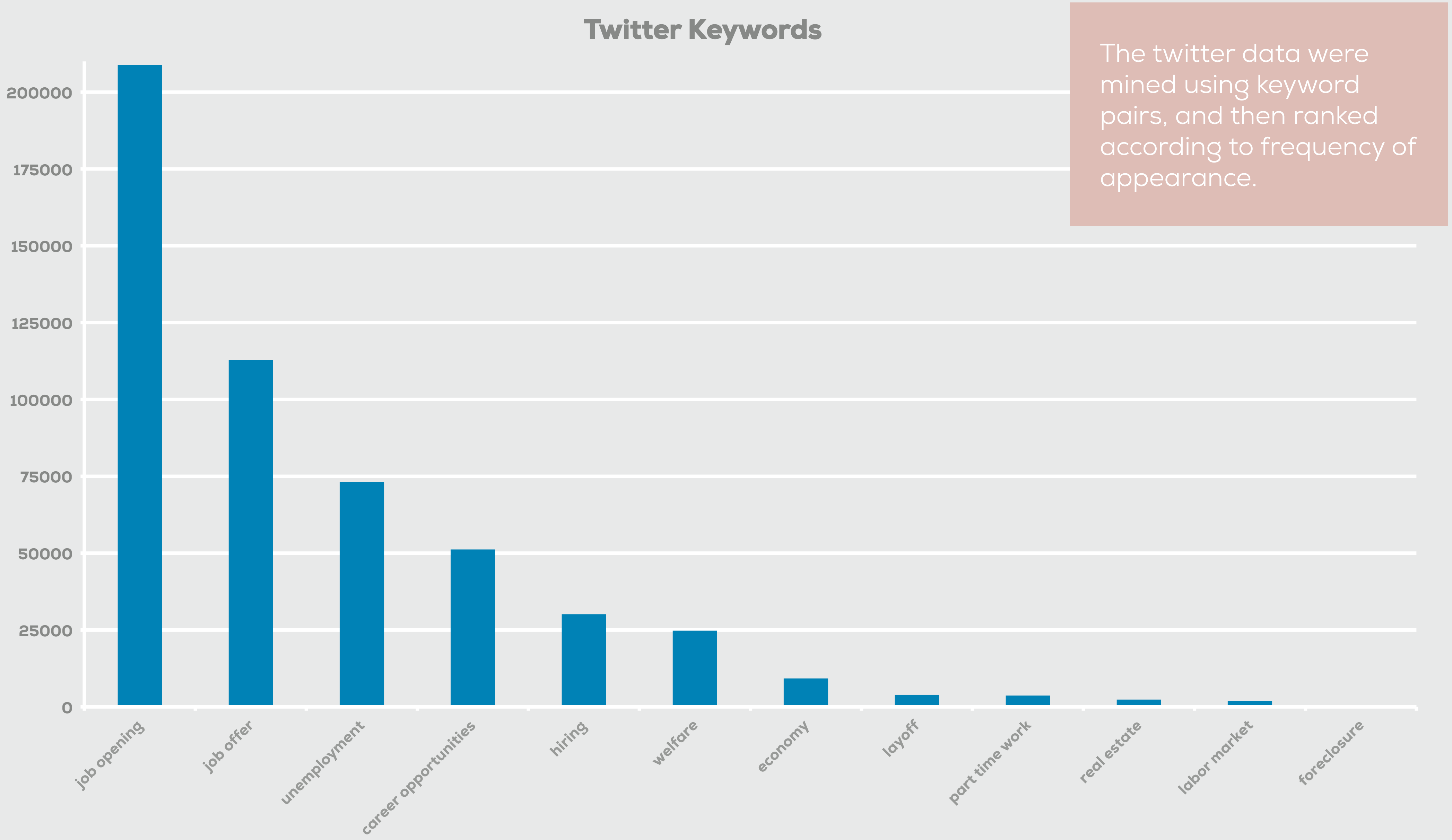
## twitter data collection

**keywords**

| | AND |
|---|---|
| job | offer |
| job | opening |
| unemployment | claim |
| work | sick |
| work | ill |
| real estate | foreclosure |
| real estate | agent |
| job | layoff |
| job | recruiting |
| social | welfare |
| food | stamp |
| house | price |
| work | hours |
| labour | market |
| hiring | usa |
| career | opportunities |
| job | wage |
| job | market |
| work | part time |
| unemployment | benefits |
| unemployment | insurance |

# PROCESS

keyword isolation



The twitter data were mined using keyword pairs, and then ranked according to frequency of appearance.

# PROCESS

## sentiment calculation

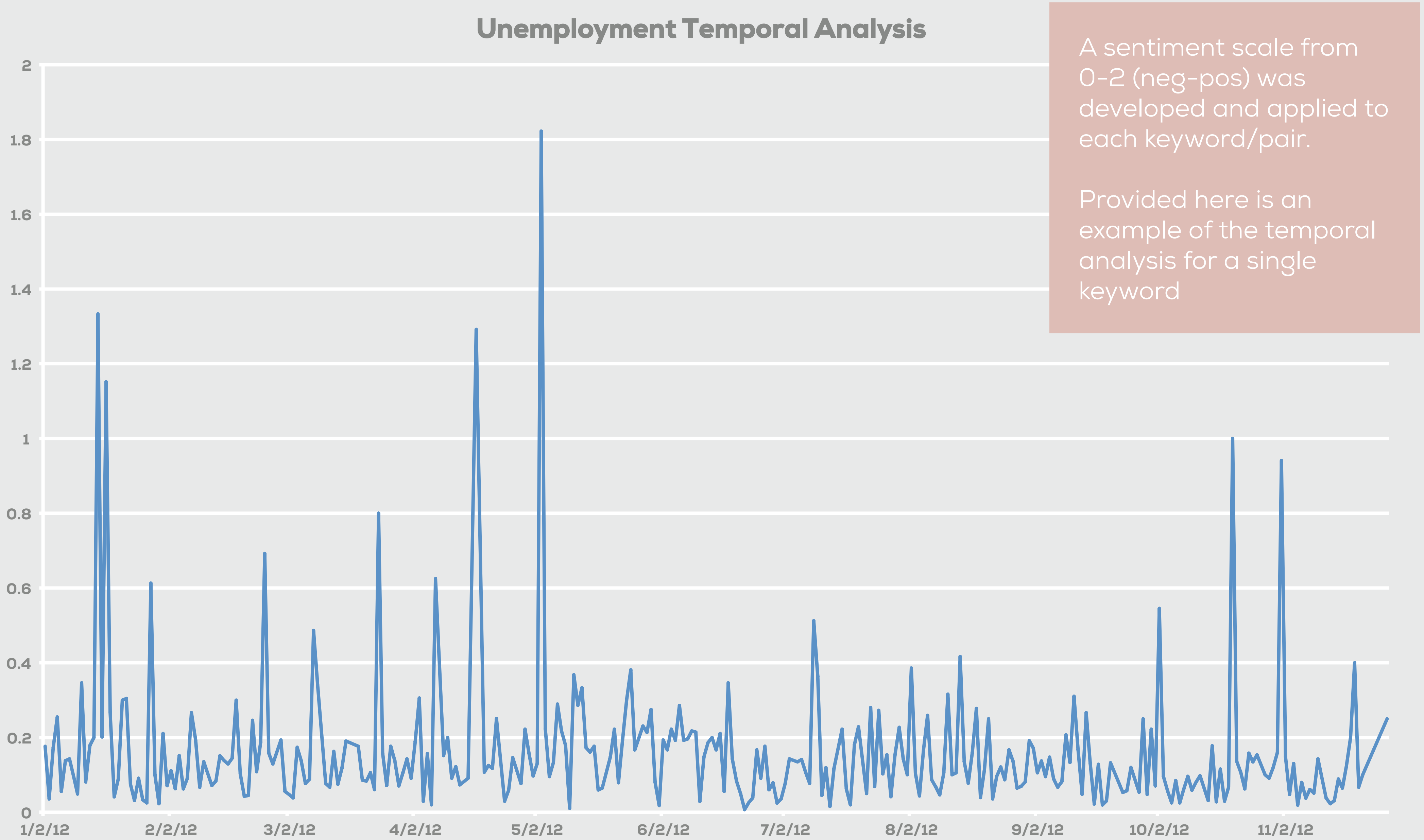
**Sentiment analysis performed by Kang Zhang using LingPipe[1]**

- LingPipe is a toolkit that processes text using computational linguistics
- LingPipe's language classification framework was used that rates text as positive, negative or neutral
- A set of a few thousand tweets that were manually ranked  was used as the training set for the classifier

1- (http://alias-i.com/lingpipe)

# PROCESS

## sentiment calculation



A sentiment scale from 0-2 (neg-pos) was developed and applied to each keyword/pair.

Provided here is an example of the temporal analysis for a single keyword

# PROCESS

## index design

Most of the important economic indicators are monthly or quarterly.

There are some high frequency weekly indicators, which don’t lead the economy, but they are a snapshot of the virtual present, as opposed to looking in the rear view mirror.

While there is plenty of noise, they should show turns or continuations in a trend ***before they show up in monthly or quarterly data.***

# PROCESS

## index design

**EmplIndex1** = JobIndex*JobWeight + LabourIndex*LabourWeight + CarrerIndex*CarrerWeight

**EmplIndex2** = JobIndex*JobWeight + LabourIndex*LabourWeight +CarrerIndex*CarrerWeight)*3 *EmotionIndex

**EmplIndex3** = JobIndex*JobWeight + 3*LabourIndex*LabourWeight + 2*CarrerIndex*CarrerWeight

**EmplIndex4** = JobIndex*3*EmotionIndex

*where:*
**Keyword Index** =#Positive keyword tweets/ #Negative keyword tweets

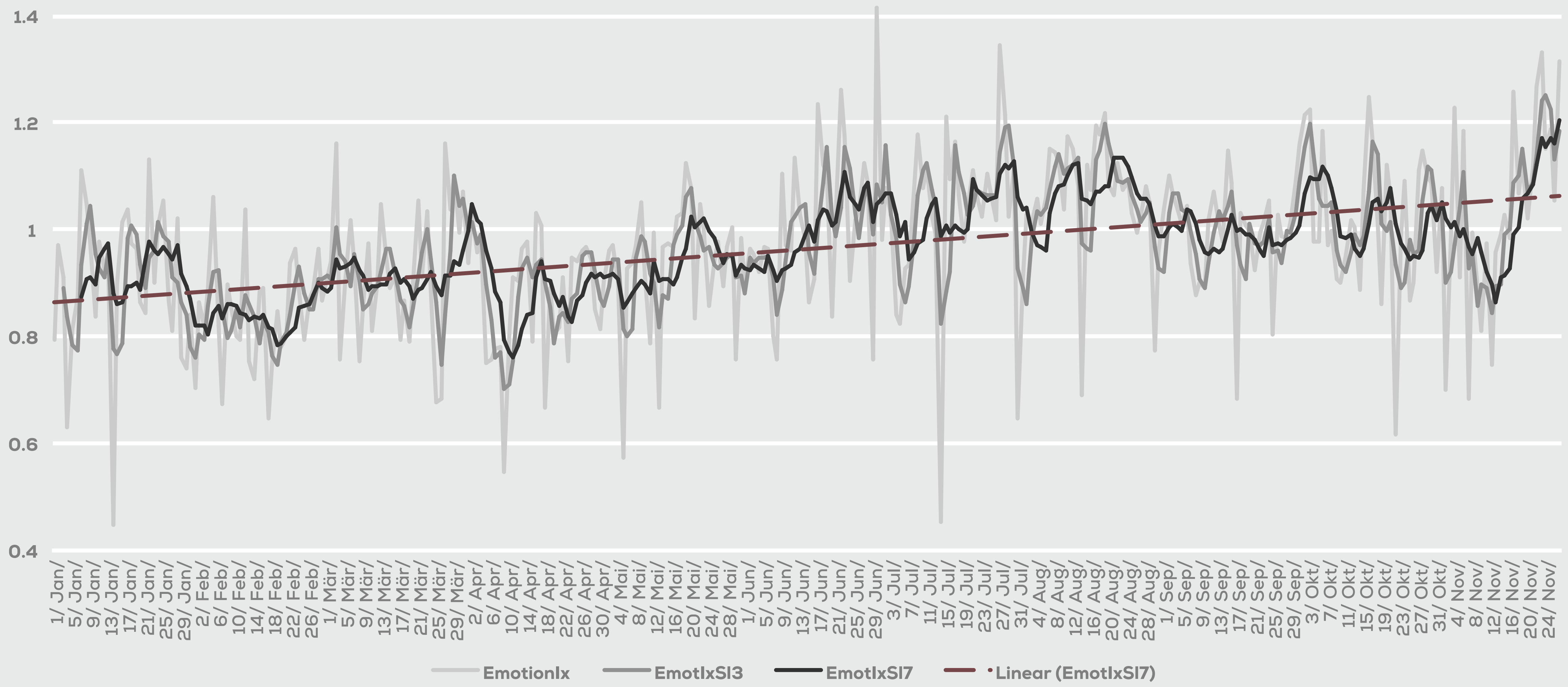
**Keyword Weight** = (#Keyword tweets) / #Total tweets number

**EmotionIndex** = (#Positive tweets / #Negative tweets) / (#Total tweets number)

# PROCESS

index
correlation

Emotional Index and different sliding averages



# PROCESS

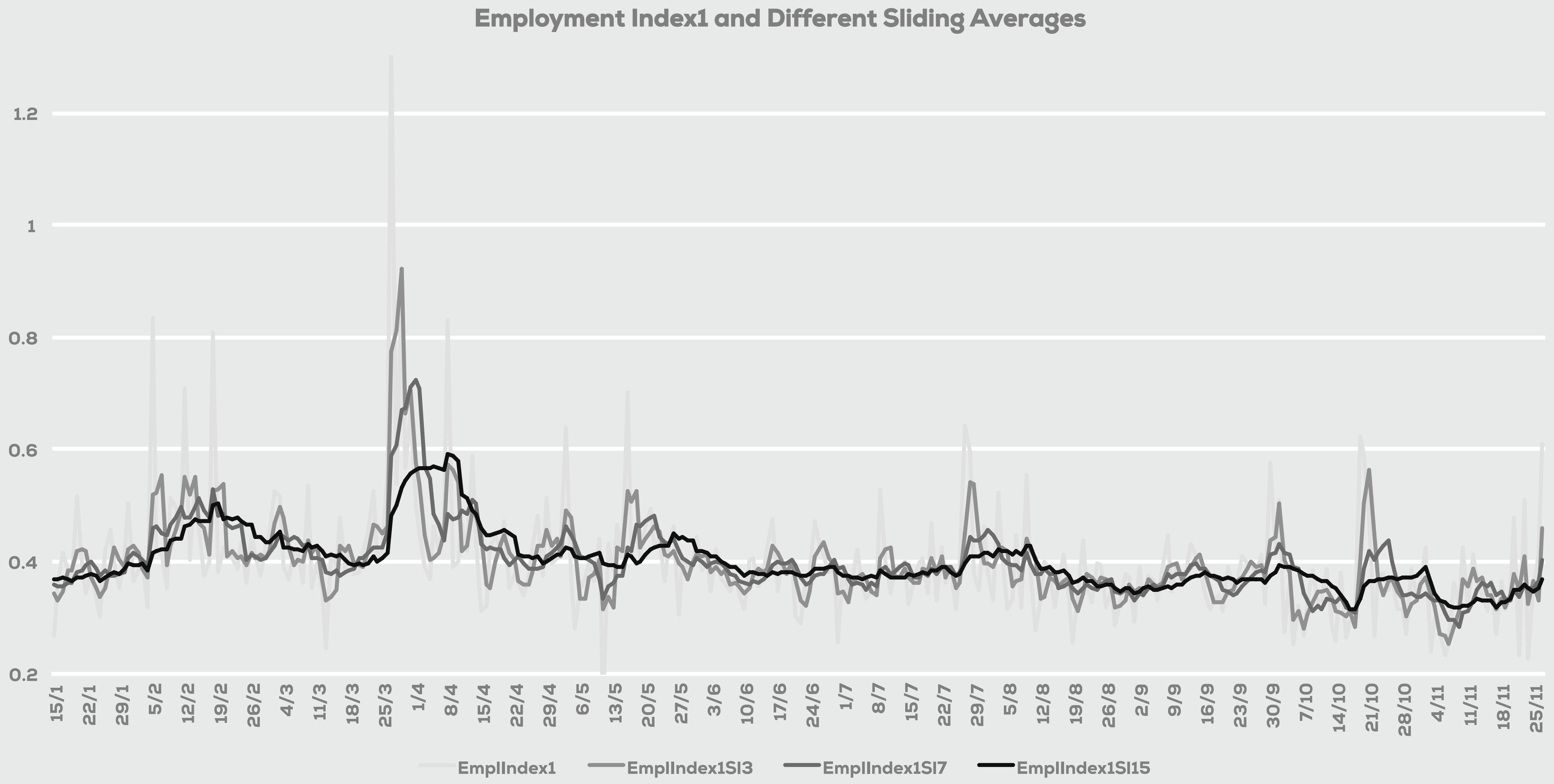
## index design

Application of different sliding averages results in 16 different Indexes/independent variables

| | *initial daily index* | *3 day sliding average* | *7 day sliding average* | *15 day sliding average* |
|---|---|---|---|---|
| index 1 | EmplIndex1 | EmplIndex1SI1 | EmplIndex1SI7 | EmplIndex1SI15 |
| index 2 | EmplIndex2 | EmplIndex2SI1 | EmplIndex2SI7 | EmplIndex2SI15 |
| index 3 | EmplIndex3 | EmplIndex3SI1 | EmplIndex3SI7 | EmplIndex3SI15 |
| index 4 | EmplIndex4 | EmplIndex4SI1 | EmplIndex4SI7 | EmplIndex4SI15 |

# PROCESS

index
correlation



# PROCESS

## index design

Application of different sliding averages results in 16 different Indexes/ independent variables

| | |
|---|---|
| var1 | Employment Level |
| var2 | (Seas) Civilian Labor Force Level |
| var3 | Civilian labor force participation rate |
| var4 | Unemployment Level |
| var5 | Unemployment rate |
| var6 | Employment-population ratio |
| var7 | Unemployment Rate - 16-19 yrs. |
| var8 | Unemployment Rate - 20 yrs. & over, Men |
| var9 | Unemployment Rate - 20 yrs. & over, Women |
| var10 | Unemployment Rate - White |
| var11 | Unemployment Rate - Black or African American |
| var12 | Unemployment Rate - Hispanic or Latino |
| var13 | Unemployment Rate - Less than a High School Diploma, 25 yrs. & over |
| var14 | Unemployment Rate - High School Graduates, No College, 25 yrs. & over |
| var15 | Unemployment Rate - Some College or Associate Degree, 25 yrs. & over |
| var16 | Unemployment Rate - Bachelor's degree and higher, 25 yrs. & over |
| var17 | Number Unemployed for Less than 5 Weeks |
| var18 | Number Unemployed for 5-14 Weeks |
| var19 | Average Weeks Unemployed |
| var20 | Unemployment Level - Job Losers |
| var21 | Unemployment Level - Reentrants to Labor Force |
| var22 | Employment Level - Part-Time for Economic Reasons, All Industries |
| var23 | Total unemployed, plus all marginally attached workers plus total employed part time for economic reasons, as a percent of all civilian labor force plus all marginally attached workers |

# PROCESS

## index design

| Variables | Expected correlation direction | Avarage correlation Direction | EmplIndex3SI3 |
|---|---|---|---|
| Employment Leve | Positiv | Negativ | -0.53 |
| (Seas) Civilian Labor Force Level | Positiv | Negativ | -0.29 |
| Civilian labor force participation rate | na | Positiv | 0.55 |
| Unemployment Level | Negative | Positiv | 0.65 |
| Unemployment rate | Negative | Positiv | 0.66 |
| Employment-population ratio | Positiv | Negativ | -0.15 |
| Unemployment Rate – 16-19 yrs. | Negative | Positiv | 0.34 |
| Unemployment Rate – 20 yrs. & over, Men | Negative | Positiv | 0.64 |
| Unemployment Rate – 20 yrs. & over, Women | Negative | Positiv | 0.47 |
| Unemployment Rate – White | Negative | Positiv | 0.67 |
| Unemployment Rate – Black or African American | Negative | not clear | 0.03 |
| Unemployment Rate – Hispanic or Latino | Negative | Positiv | 0.79 |
| Unemployment Rate – Less than a High School Diploma, 25 yrs. & over | Negative | Positiv | 0.7 |
| Unemployment Rate – High School Graduates, No College, 25 yrs. & over | Negative | Negativ | -0.27 |
| Unemployment Rate – Some College or Associate Degree, 25 yrs. & over | Negative | Positiv | 0.82 |
| Unemployment Rate – Bachelor?s degree and higher, 25 yrs. & over | Negative | Negativ | 0.12 |
| Number Unemployed for Less than 5 Weeks | Negative | Negativ | -0.17 |
| Number Unemployed for 5-14 Weeks | Negative | Positiv | 0.36 |
| Average Weeks Unemployed | Negative | Negativ | -0.11 |
| Unemployment Level – Job Losers | Negative | Positiv | 0.55 |
| Unemployment Level – Reentrants to Labor Force | Negative | Positiv | 0.58 |
| Employment Level – Part-Time for Economic Reasons, All Industrie | na | Negativ | -0.28 |
| *Total unemployed, plus all marginally attached workers plus total employed part time for economic reasons* | Negative | Negativ | 0.45 |

# RESULTS

## index correlation

Correlation of the selected variables and daily indexes

| | Correlation | P-value | | | |
|---|---|---|---|---|---|
| | *EmplIndex1SI3* | *EmplIndex3SI3* | *EmplIndex1SI3* | *EmplIndex3SI3* | |
| var5 | 0.64 | 0.66 | 0.0446 | 0.0385 | Unemployment rate |
| var12 | 0.84 | 0.79 | 0.0021 | 0.007 | Unemployment Rate - Hispanic or Latino |
| var13 | 0.72 | 0.7 | 0.0196 | 0.0238 | Unemployment Rate - Less than a High School Diploma, 25 yrs. & over |
| var15 | 0.82 | 0.82 | 0.0037 | 0.0036 | Unemployment Rate - Some College or Associate Degree, 25 yrs. & over |

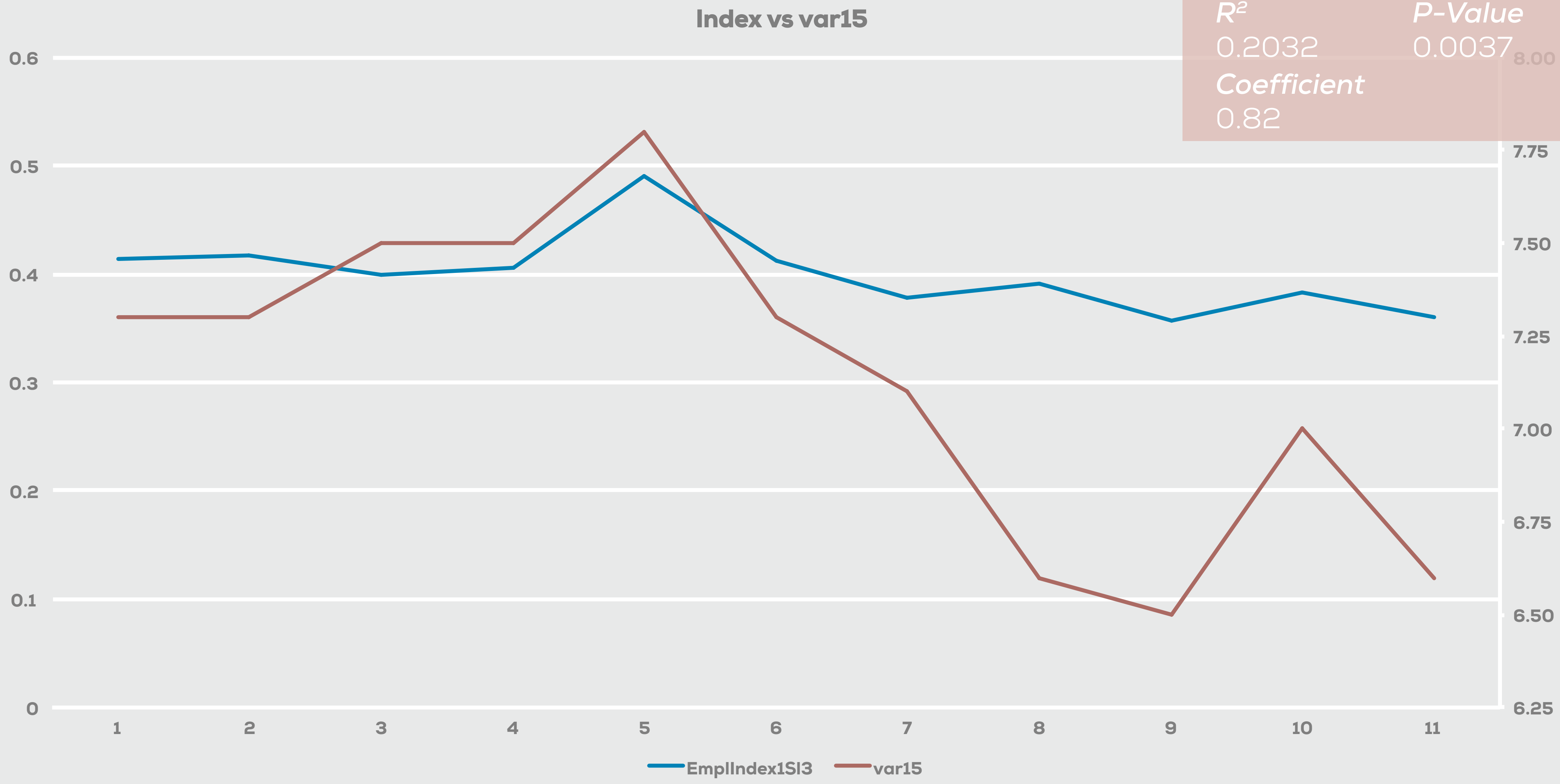
# RESULTS

## index correlation

| *R²* | *P-Value* |
|---|---|
| 0.2032 | 0.0037 |

*Coefficient*

0.82



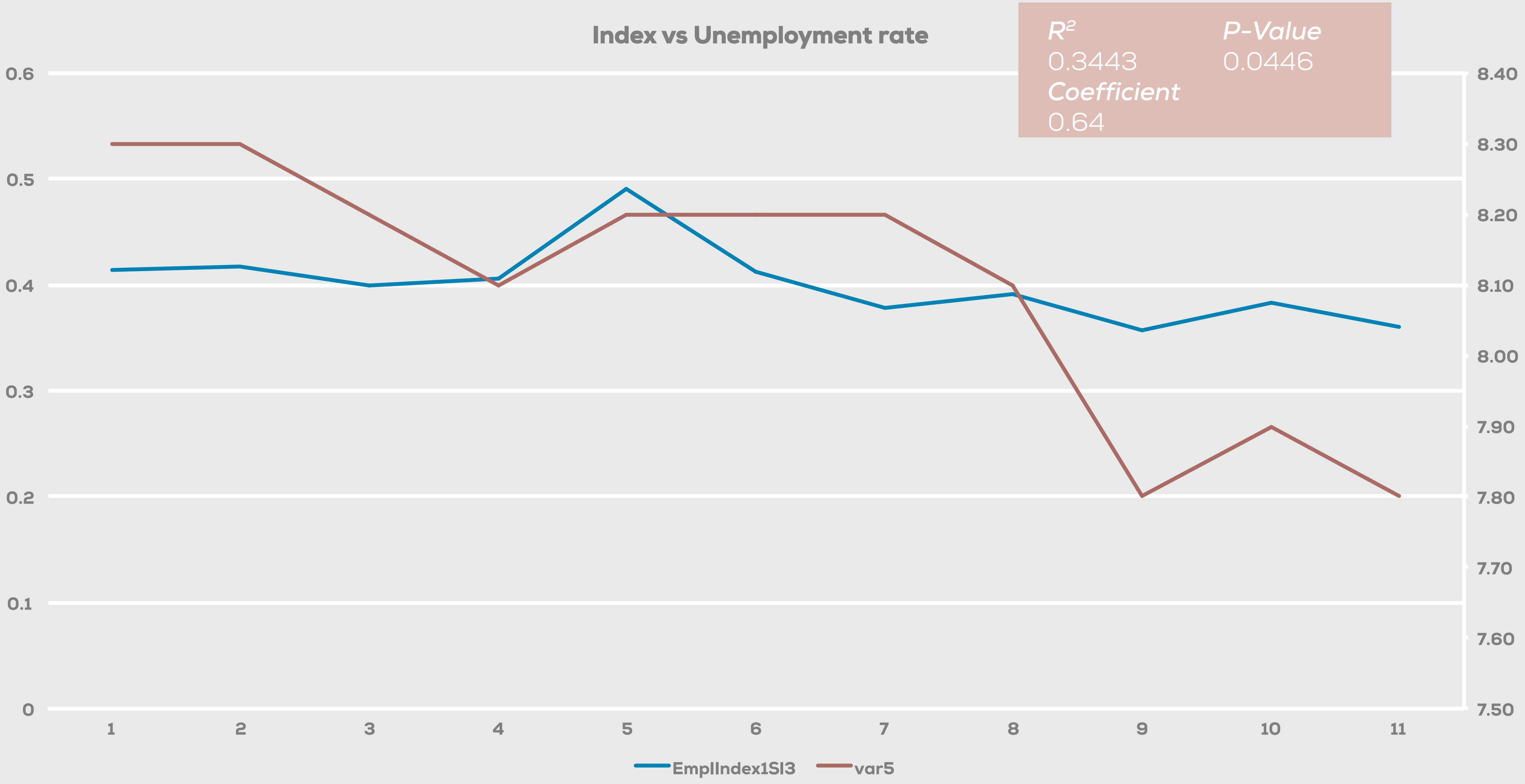
# RESULTS

index correlation

*R²*
0.3443

*P-Value*
0.0446

*Coefficient*
0.64



# RESULTS

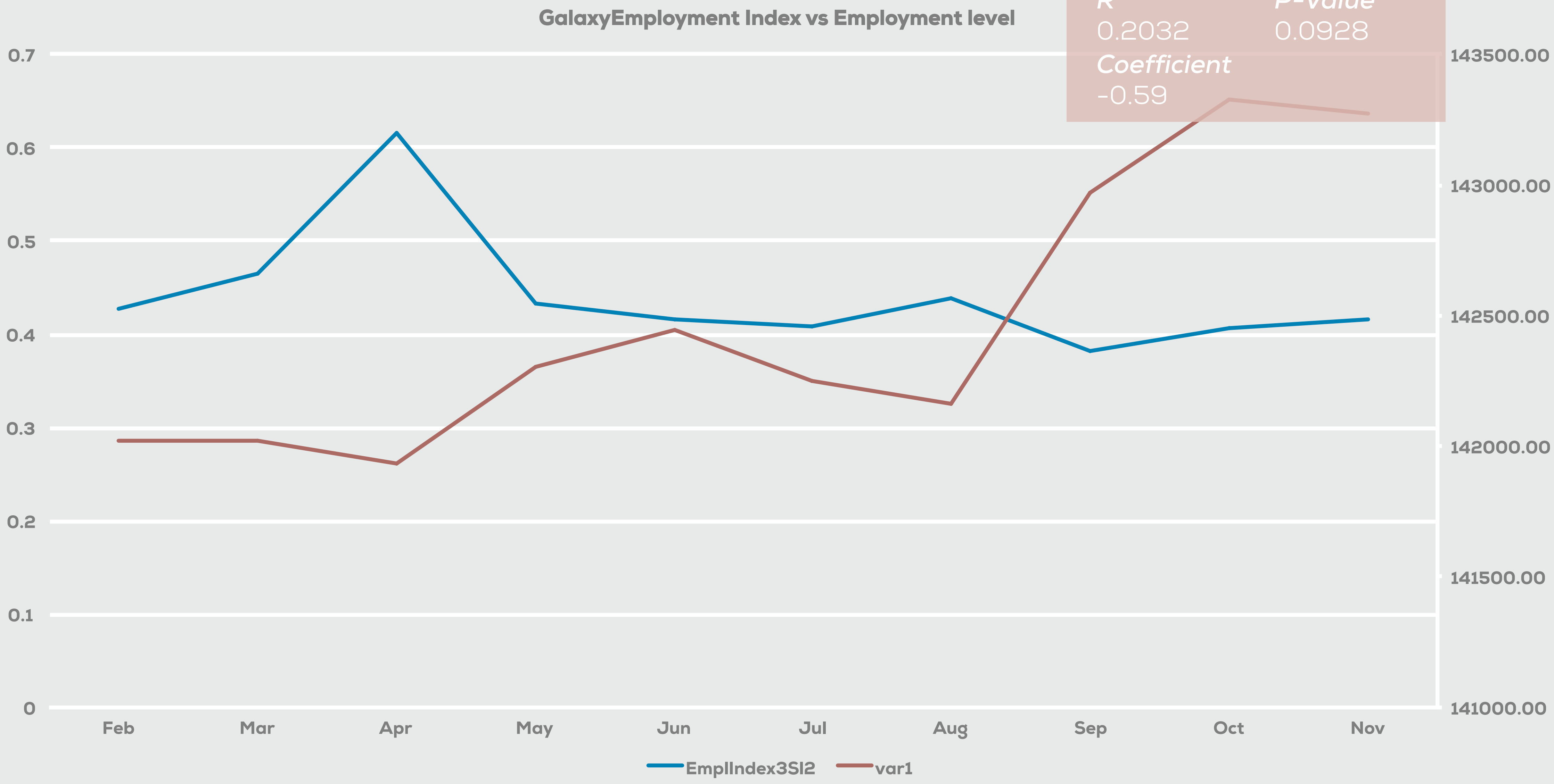
weekly index

Weekly Index is calculated as a weekly average of daily Indexes and by applying a two day sliding average.

| | Correlation | P-values | | | |
|---|---|---|---|---|---|
| | *EmplIndex1SL2* | *EmplIndex3Sl2* | *EmplIndex1SL2* | *EmplIndex3Sl2* | |
| var1 | -0.64 | -0.59 | 0.0642 | 0.0928 | - Employment Level |
| var2 | -0.76 | -0.71 | 0.0181 | 0.0323 | - (Seas) Civilian Labor Force Level |
| var22 | 0.76 | 0.73 | 0.0432 | 0.0485 | - Employment Level - Part-Time for Economic Reasons, All Industrie |
| var7 | -0.68 | -0.67 | 0.0176 | 0.0246 | - Unemployment Rate - 16-19 yrs. |

# RESULTS

weekly index



*$R^2$*
0.2032

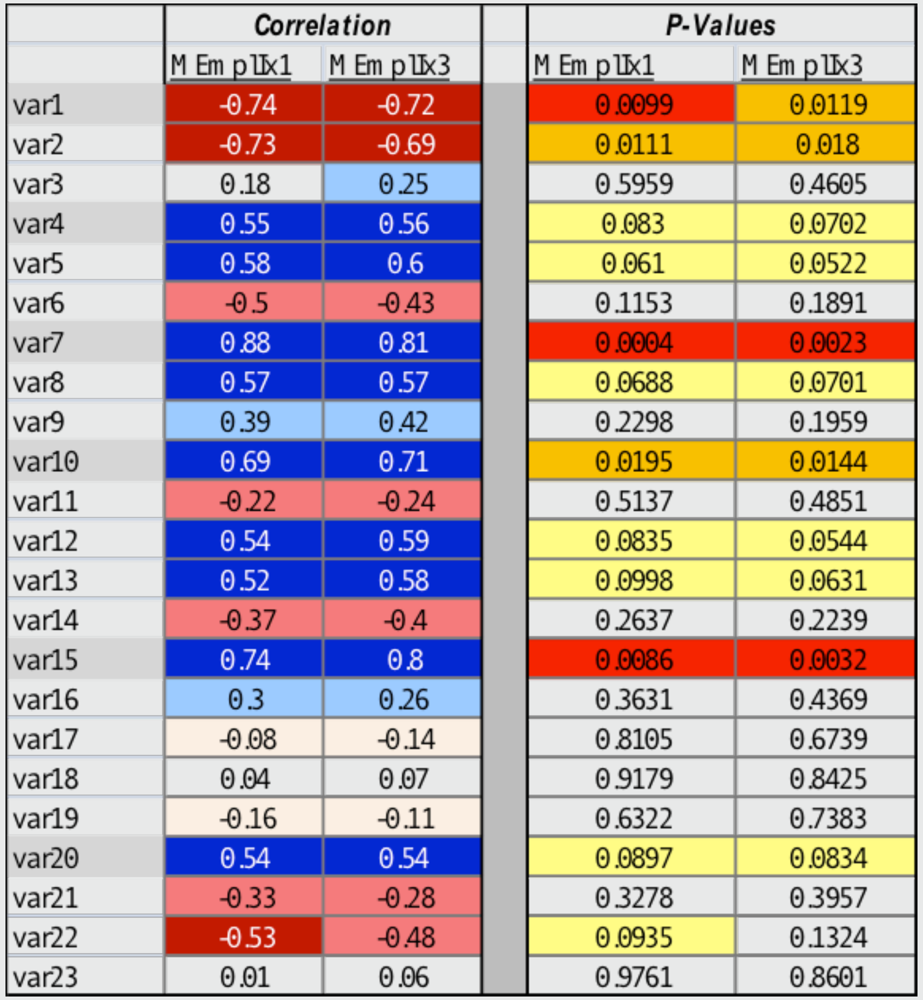
*P-Value*
0.0928

*Coefficient*
-0.59

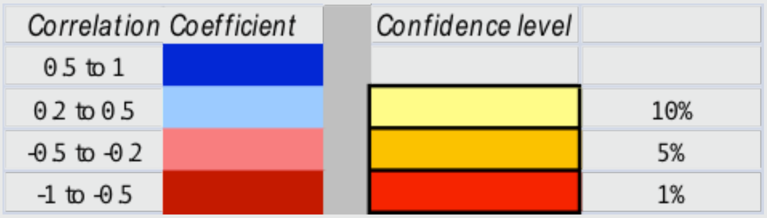
# RESULTS

## monthly indexes

Monthly Index is calculated as a monthly average of daily Indexes



| | Correlation | | P-Values | |
|---|---|---|---|---|
| | M Em plIx1 | M Em plIx3 | M Em plIx1 | M Em plIx3 |
| var1 | -0.74 | -0.72 | 0.0099 | 0.0119 |
| var2 | -0.73 | -0.69 | 0.0111 | 0.018 |
| var3 | 0.18 | 0.25 | 0.5959 | 0.4605 |
| var4 | 0.55 | 0.56 | 0.083 | 0.0702 |
| var5 | 0.58 | 0.6 | 0.061 | 0.0522 |
| var6 | -0.5 | -0.43 | 0.1153 | 0.1891 |
| var7 | 0.88 | 0.81 | 0.0004 | 0.0023 |
| var8 | 0.57 | 0.57 | 0.0688 | 0.0701 |
| var9 | 0.39 | 0.42 | 0.2298 | 0.1959 |
| var10 | 0.69 | 0.71 | 0.0195 | 0.0144 |
| var11 | -0.22 | -0.24 | 0.5137 | 0.4851 |
| var12 | 0.54 | 0.59 | 0.0835 | 0.0544 |
| var13 | 0.52 | 0.58 | 0.0998 | 0.0631 |
| var14 | -0.37 | -0.4 | 0.2637 | 0.2239 |
| var15 | 0.74 | 0.8 | 0.0086 | 0.0032 |
| var16 | 0.3 | 0.26 | 0.3631 | 0.4369 |
| var17 | -0.08 | -0.14 | 0.8105 | 0.6739 |
| var18 | 0.04 | 0.07 | 0.9179 | 0.8425 |
| var19 | -0.16 | -0.11 | 0.6322 | 0.7383 |
| var20 | 0.54 | 0.54 | 0.0897 | 0.0834 |
| var21 | -0.33 | -0.28 | 0.3278 | 0.3957 |
| var22 | -0.53 | -0.48 | 0.0935 | 0.1324 |
| var23 | 0.01 | 0.06 | 0.9761 | 0.8601 |



| Correlation Coefficient | | Confidence level | |
|---|---|---|---|
| 0 5 to 1 | 🟦 | | |
| 0 2 to 0 5 | 🩵 | 🟨 | 10% |
| -0 5 to -0 2 | 🩷 | 🟧 | 5% |
| -1 to -0 5 | 🟥 | 🟥 | 1% |

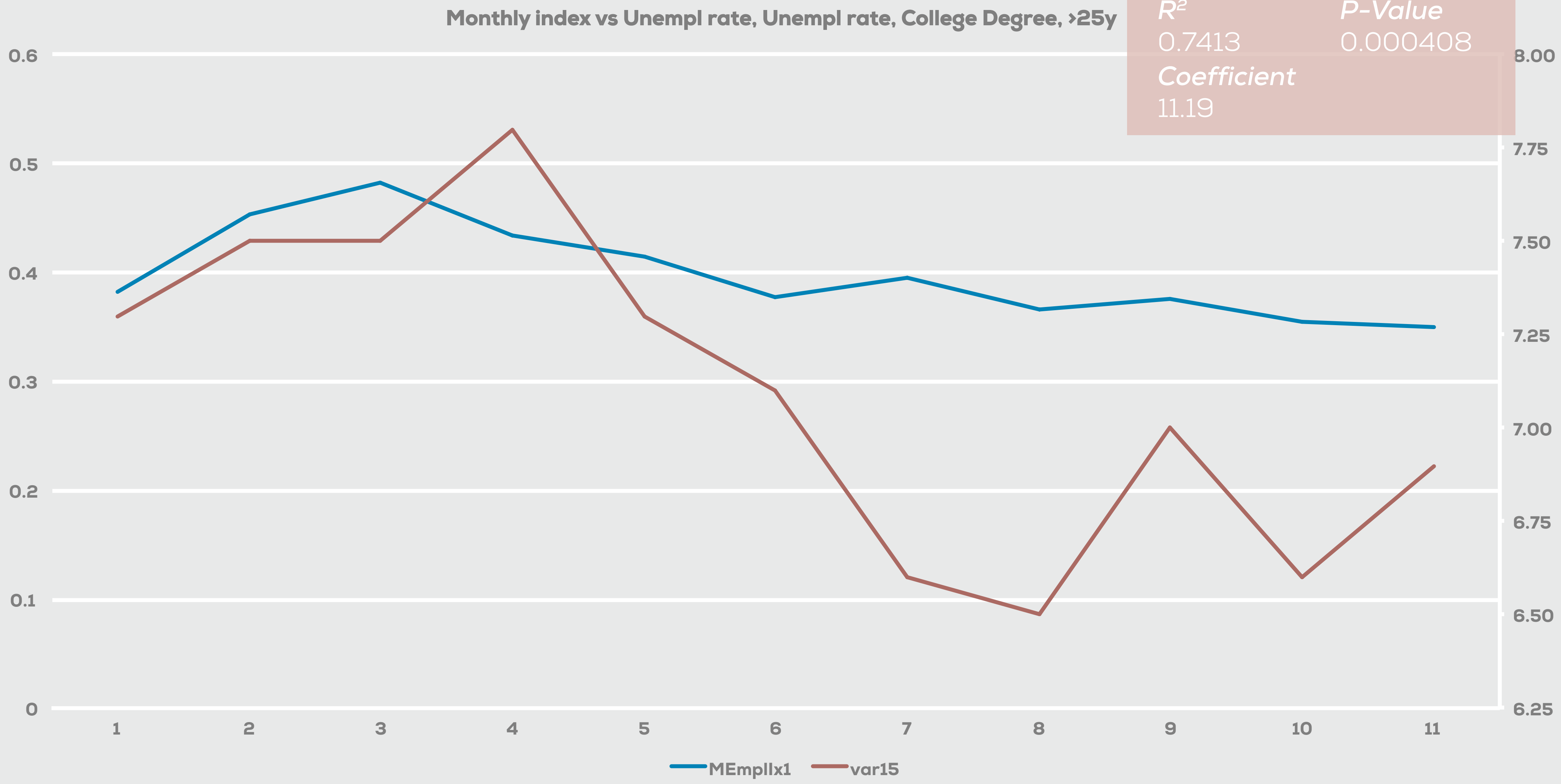
# RESULTS

## monthly index

| Regression | Employment Leve | | Unemployment rate | | Unempl rate, College Degree, ›25 | | Unempl. Rate - 16-19 yrs. | |
|---|---|---|---|---|---|---|---|---|
| | *MEmplIx1* | | *MEmplIx1* | | *MEmplIx3* | | *MEmplIx1* | |
| | | | | | | | | |
| **Adjusted R-squared** | 0.4893 | | 0.2637 | | 0.5963 | | 0.7413 | |
| **coefficient** | | P - value | | P - value | | P - value | | P - value |
| **(Intercept)** | 146400 | 7.89E-16 | 7.013 | 6.59E-14 | 3.6514 | 2.35E-03 | 19.5901 | 2.35E-03 |
| **Index** | -9657 | 0.00995** | 2.6109 | 0.061. | 7.8216 | 0.00325** | 11.1914 | 0.000408*** |
| | | | | | | | | |
| **F-statistic** | 10.58 | 9DF | 4.582 | 9DF | 15.77 | 9DF | 15.77 | 9DF |

# RESULTS

## monthly index



*$R^2$*
0.7413

*P-Value*
0.000408

*Coefficient*
11.19

# RESULTS

## Daily Galaxy Employment Index and Financial market movements

financial index

| | Correlation | | | | P - Values | | | |
|---|---|---|---|---|---|---|---|---|
| | *SPCOMP* | *SPCOMPsI3* | *SPCOMPsI7* | *SPCOMPsI15* | *SPCOMP* | *SPCOMPsI3* | *SPCOMPsI7* | *SPCOMPsI15* |
| **EmplIndex1** | -0.08 | -0.08 | -0.08 | -0.11 | 0.247 | 0.2428 | 0.2169 | 0.0923 |
| **EmplIndex1SI3** | -0.16 | -0.15 | -0.14 | -0.17 | 0.0155 | 0.0268 | 0.0319 | 0.0117 |
| **EmplIndex1SI7** | -0.16 | -0.16 | -0.16 | -0.19 | 0.0166 | 0.0159 | 0.0169 | 0.005 |
| **EmplIndex1SI15** | -0.18 | -0.17 | -0.17 | -0.19 | 0.007 | 0.0098 | 0.0113 | 0.0043 |
| **EmplIndex2** | 0.02 | 0.01 | 0.01 | -0.02 | 0.7224 | 0.8597 | 0.9072 | 0.7801 |
| **EmplIndex2SI3** | -0.01 | -0.02 | -0.03 | -0.05 | 0.8463 | 0.7899 | 0.6659 | 0.4592 |
| **EmplIndex2SI7** | 0.03 | 0.02 | -0.01 | -0.04 | 0.7028 | 0.8176 | 0.9372 | 0.5687 |
| **EmplIndex2SI15** | 0.02 | 0.03 | 0.03 | 0 | 0.7651 | 0.6734 | 0.6854 | 0.9413 |

# RESULTS

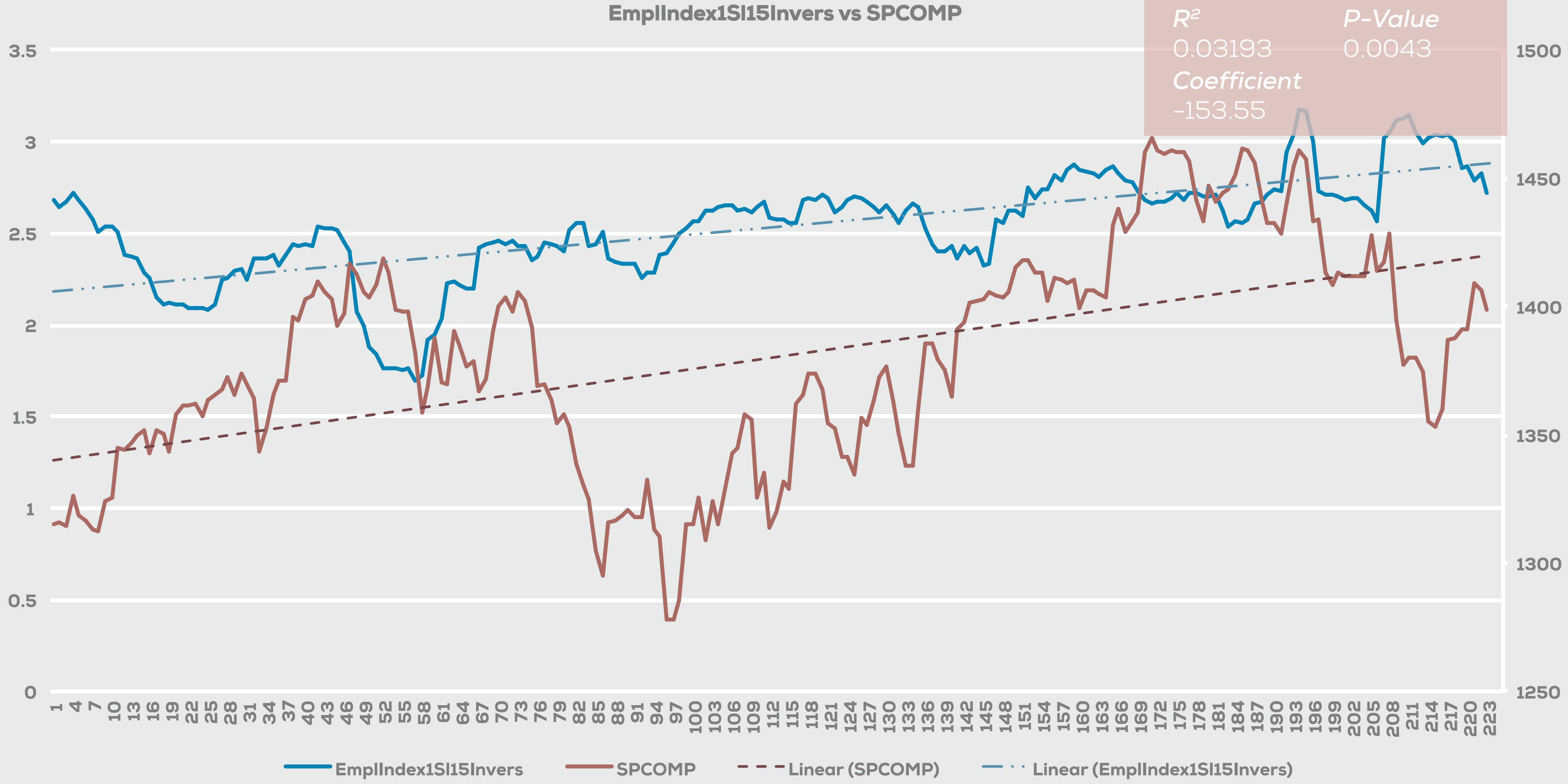
financial index

## Daily Galaxy Employment Index and Financial market movements

| Regression | S&PCOMP with | | | | S&PCOMP 15 days sliding avg | |
|---|---|---|---|---|---|---|
| | *EmplIndex1SI15* | | *EmplIndex1SI7* | | | |
| **Adjusted R-squared** | *0.02801* | | *0.0213* | | *0.03193* | |
| **coefficient** | | P - value | | P - value | | P - value |
| **(Intercept)** | 1440.54 | 2E-16 | 1.43E+03 | 2.00E-16 | 1439.31 | 2.00E-16 |
| **Index** | -151.05 | 0.007047 | -113.6 | 0.0166 | -153.55 | 0.0043 |
| **F-statistic** | 7.398 | 221DF | 5.831 | 221DF | 8.323 | 221DF |

# RESULTS

index correlation



| *$R^2$* | *P-Value* |
|---|---|
| 0.03193 | 0.0043 |

*Coefficient*

-153.55

# RESULTS

## weekly index

| | Correlation Coeficient | | P - Value | |
|---|---|---|---|---|
| | DJINDUS | S.PCOMP | DJINDUS | S.PCOMP |
| **EmotIx** | -0.24 | -0.13 | 0.2149 | 0.4986 |
| **EmplIndex1** | 0.35 | 0.39 | 0.0606 | 0.0387 |
| **EmplIndex1SL2** | 0.46 | 0.51 | 0.0118 | 0.0045 |
| **EmplIndex2** | 0.24 | 0.32 | 0.2078 | 0.0914 |
| **EmplIndex2Sl2** | 0.33 | 0.45 | 0.0819 | 0.0148 |

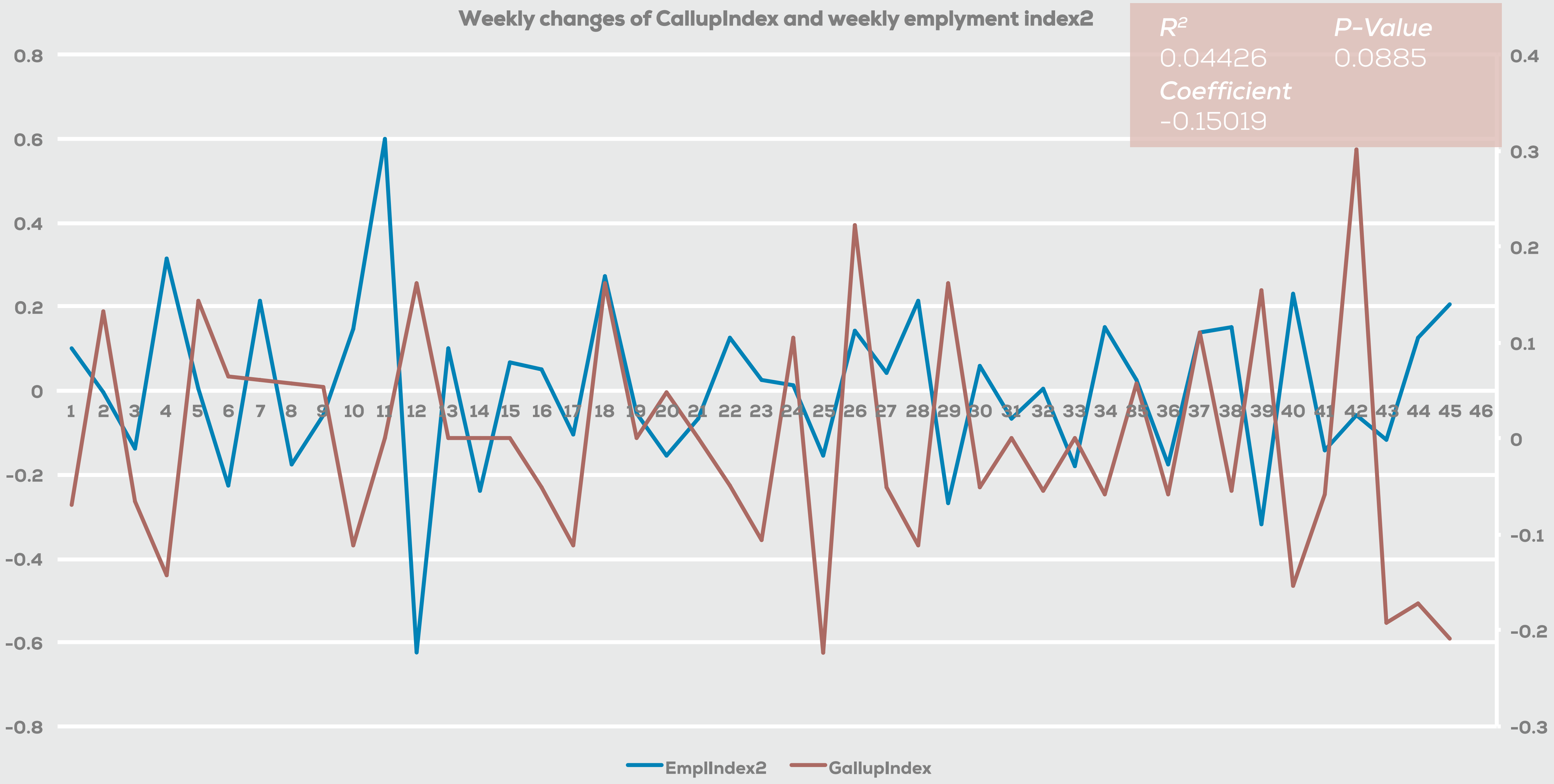
*First 30 weeks of 2012: Galaxy Index vs. Financial Markets*

| **Regression** | **S&PCOMP weekly avg** | |
|---|---|---|
| | Weekly EmplIndex1Sl2 | |
| **Adjusted R-squared** | 0.235 | |
| **coefficient** | | P – value |
| **(Intercept)** | 1220.38 | 2.00E-16 |
| **Index** | 313.46 | 0.0045 |
| **F-statistic** | 9.603 | 27DF |

| **Regression** (two independent variables) | **S&PCOMP weekly avg** | |
|---|---|---|
| | 1. Weekly EmplIndex1Sl2<br>2. Emotion Index | |
| **Adjusted R-squared** | 0.235 | |
| **coefficient** | | P - value |
| **(Intercept)** | 1194.03 | 2.00E-16 |
| **EmotIx** | 22.77 | 0.75 |
| **EmplIndex1SL2** | 325.53 | 0.00645 |
| **F-statistic** | 9.603 | 27DF |

# RESULTS

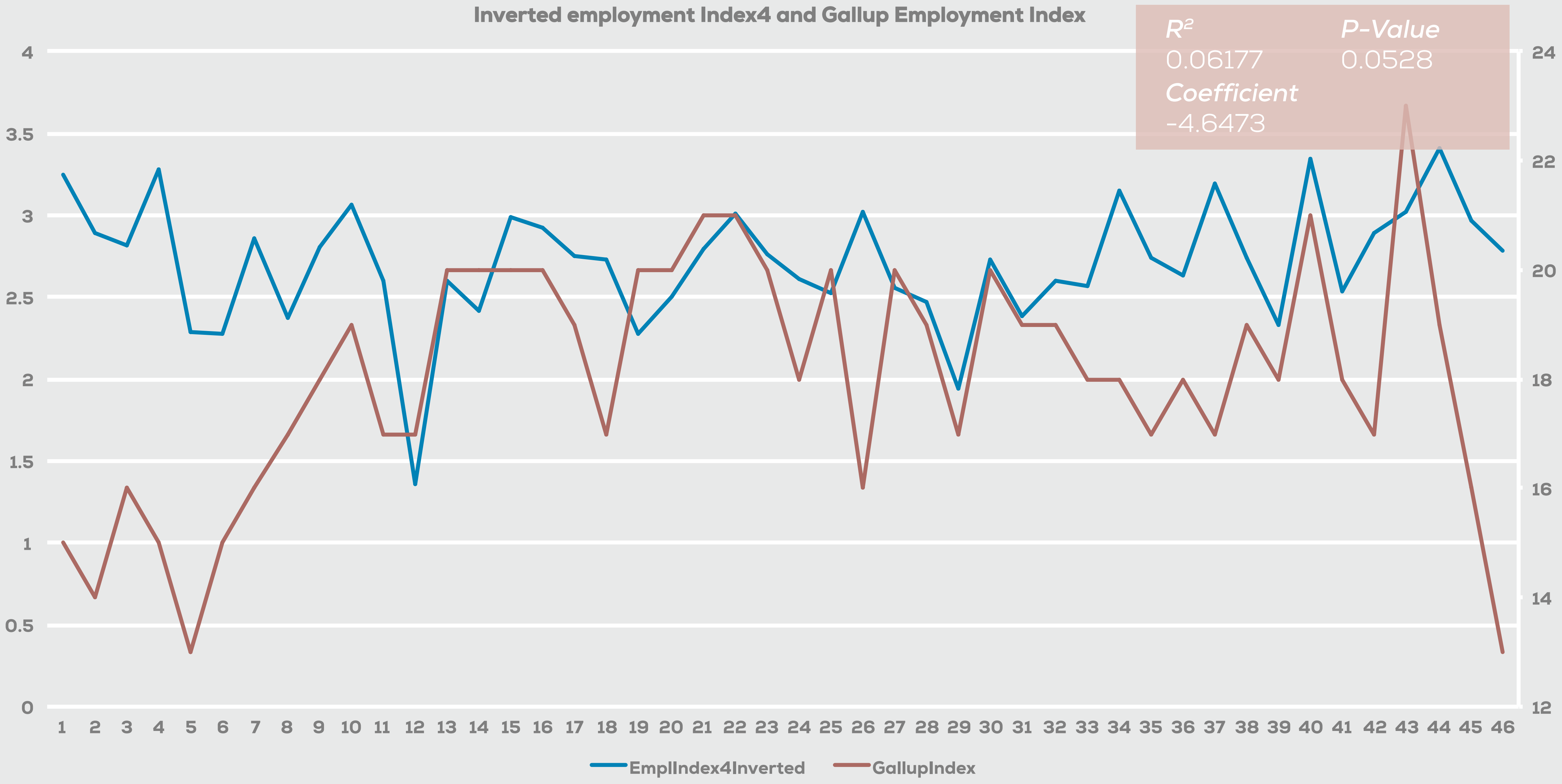
index correlation



| *$R^2$* | *P-Value* |
|---|---|
| 0.04426 | 0.0885 |
| *Coefficient* | |
| -0.15019 | |

# RESULTS

## index correlation



Inverted employment Index4 and Gallup Employment Index

| *$R^2$* | *P-Value* |
|---|---|
| 0.06177 | 0.0528 |
| *Coefficient* | |
| -4.6473 | |

# RESULTS SUMMARY

- In our project we estimated that sentiment analysis of the employment situation based on Twitter data replicate job surveys and statistics
- We calculated daily, weekly and monthly Employment Indexes
- In order to smooth the high volatility of daily indexes we applied different  sliding averages

# RESULTS SUMMARY

- The Daily Employment Indexes 1 and 3 with 3 days sliding average shows high correlation with some monthly unemployment rates.
- The weekly Indexes 1 and 3 with 2 weeks sliding average and monthly Indexes 1 and 3 are good predictor of different employment indicators
- We also estimated correlation between financial market movements and our indexes
- The daily employment index 2 and 4 are correlated with the Gallup employment indexes

# RESULTS SUMMARY

conclusion

The expensive and time-intensive polling and surveys can be supplemented or extended by the automated analysis of the simple to gather social media data

# FURTHER

research and improvements

Keywords research by applying linguistic analysis

Improved Sentiment analysis by:

- Training sentiment algorithm on twitter data set related to the employment topic
- Extending the sentiment scale (e.g. very positive and positive)
- Using demographic and geographic information

Creating specific employment indexes based on demographic and geographic data

Twitter dataset for at least 2 years

# THANKS