**[REV.00]**

**Mahalanobis 거리를 이용한**

**Outlier 제거 방법**

**Contents**

[1 Purpose 4](#_Toc492167057)

[2 Concept 4](#_Toc492167058)

[*3* Theory *.* 8](#_Toc492167059)

[4 Test Result 10](#_Toc492167060)

[5 Expected benefits 15](#_Toc492167061)

# Purpose

조기경보 시스템은 과거에 정상적인 데이터를 습득 한 모델을 기반으로 학습되지 않은 데이터 세트를 받으면 알람을 발생시킨다. 이는 학습된 범위를 초과하는 값이 수신 될 때 경보를 발생시킬 뿐만 아니라, 학습 된 범위에 포함되는 경우에도 학습 된 패턴과 다른 패턴의 데이터 세트가 수신 될 때 경보를 발생시킨다. 따라서 장비의 잠재적 고장을 감지한다. 이는 고객에게 우리의 가치를 전하는 방법이다.

우리는 적절한 예측 값을 계산하는 것이 분명 분명히 중요하지만 고객 사이트에서 잠재적 인 실패를 찾는 것이 가장 중요하다는 것을 인식해야 한다. 그러나 학습 모델에는 가장 큰 문제가 있다. 그것은 과잉 학습이다(over-fitting). AAKR 알고리즘은 많은 샘플을 학습이 필요하기 때문에 잘못된 정보를 정상적인 정보로 인식하여 학습할 여지가 많다. 잘못된 정보(비정상 패턴)가 학습되면, 조기경보 시스템은 고객의 잠재적 문제를 발견하지 못할 수도 있다.

따라서 올바른 예측과 조기 경보를 위해 비정상 데이터(Outlier)를 학습 모델에서 제거해야 한다.

# Concept

샘플 데이터에 이상치가 있는 경우 샘플은 다른 샘플 데이터와 멀리 떨어져 있어야한다. 따라서 Mahalanobis distance를 사용하여 어떤 샘플 데이터가 이상치 인지 알 수 있고 각 샘플에서 평균 샘플까지의 Mahalanobis distance를계산할 수 있다. 또한 Mahalanobis distance가 먼 샘플들을 제거 할 수 있다.

\*

Mahalanobis distance는 1936 년 P. C. Mahalanobis에 의해 소개된 P 점과 D 점 사이의 거리를 측정 한 값이다. [1] 이것은 P가 D의 평균으로부터 몇 표준 편차만큼 떨어져 있는지를 측정하는 개념을 다차원 적으로 일반화한 것이다. P가 D의 평균에 있으면 이 거리는 0이고 P가 평균에서 멀어 질수록 커진다. 주축 구성 요소 축의 경우 P에서 D의 평균까지의 표준 편차 수를 측정한다. 이러한 각 축의 단위 분산을 조정한 경우 Mahalanobis distance는 변형된 공간의 표준 유클리드 거리에 해당한다. 따라서 Mahalanobis distance는 단위가 없고 스케일에 영향을 받지 않으며 데이터 집합의 상관 관계를 고려한다.

*출처 : 위키피디아*

그렇다면 Euclidean distance가 아닌 Mahalanobis distance를 사용하는 이유는 무엇일까?

아래 예제를 보자.

Figure 1. Line Chart - Original Data

위의 차트에서 비정상 데이터를 찾을 수 있는가? 아마 찾기 힘들 것이다.

그렇다면 아래 분산형 차트를 한번 살펴보자

**Outliers??**

Figure 2. Scatter Chart – Original Data

아마도 비정상 데이터로 의심되는 샘플을 찾을 수 있을 것이다. 분산 형 차트에서 특이점을 찾는 것이 선 차트보다 쉽다.

그러나 이 예는 그룹에 단지 두 개의 태그만 있는 경우일 뿐이며, 만약 세 개 이상의 태그가 있는 경우 분산형 차트에서도 비정상 데이터로 의심되는 샘플을 찾기 힘들 것이다.

각 샘플이 비정상 데이터인지 아닌지를 구별하기 위해 각 샘플과 전체 샘플의 평균 사이의 거리를 계산한다. Euclidian distance를 계산했을 때와 Mahalanobis distance를 계산했을 때의 차이를 살펴보자.

**B (479.9, 468.2)**

**A (446.7, 460.0)**

**Mean**

**M (453.2, 448.7)**

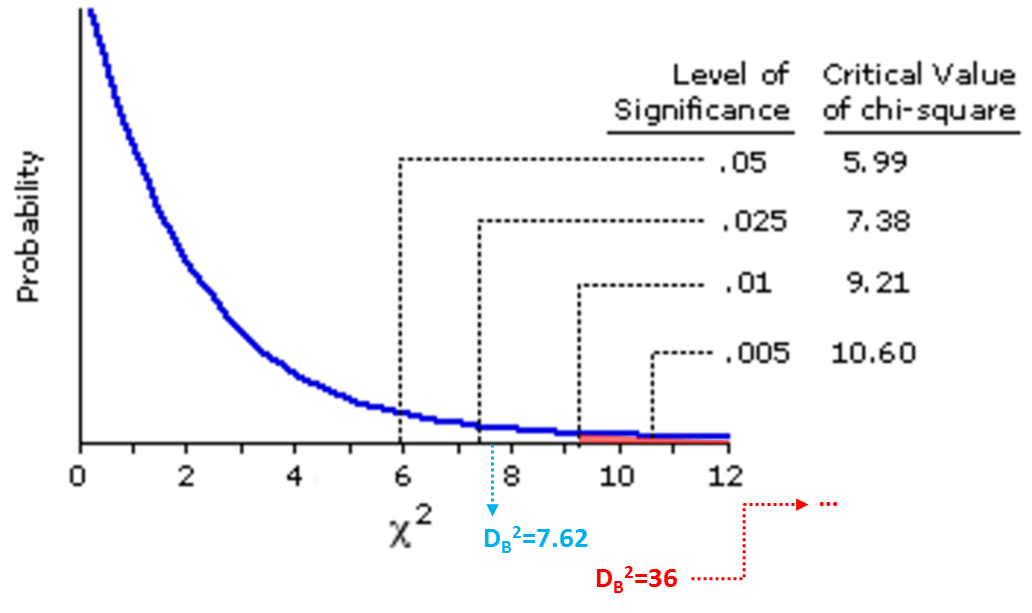
Figure 3. Scatter Chart - Distance

Euclidian distance계산 결과 B의 거리가 A보다 크다.

그러면 Mahalanobis distance를 계산하여 위 계산 결과와 비교해보자.

Euclidian distance로 계산할 때 B는 A보다 크지 만 Mahalanobis distance는 표준 편차를 고려하므로 Mahalanobis distance로 계산할 때 A의 거리는 B보다 큰 것으로 계산된다.

데이터가 k 변수를 갖는 정규 분포라고 가정하면, Mahalanobis distance square는 k 자유도를 갖는 카이 제곱을 따른다. 2 개의 변수가 있는 데이터는 아래 그림처럼 2 자유도의 카이 제곱 분포를 따르게 된다.



따라서 유의 수준을 1 % (0.01)로 설정하면 임계 값은 9.21이다. 위의 예에서 A의 카이 제곱 값은 임계 값 = 36 (62)을 초과하므로 이상 값으로 간주하여 제거할 수 있다. B의 카이 제곱 값은 임계 값 = 7.62 (2.672)를 초과하지 않으므로 제거할 필요가 없다.

Chi-Square Distribution Table 
0 c2 
The shaded area is equal to ® for Â2 = Â2 
®. 
df Â2: 
995 Â2 :990 Â2: 
975 Â2 :950 Â...

Figure 4. Chi-square distribution table

# Theory *이 섹션은 이해를 돕기 위한 페이지 일 뿐이므로 읽지 않아도 상관 없습니다.*

단 변량에 대한 정규 분포의 확률 밀도는 다음과 같다.

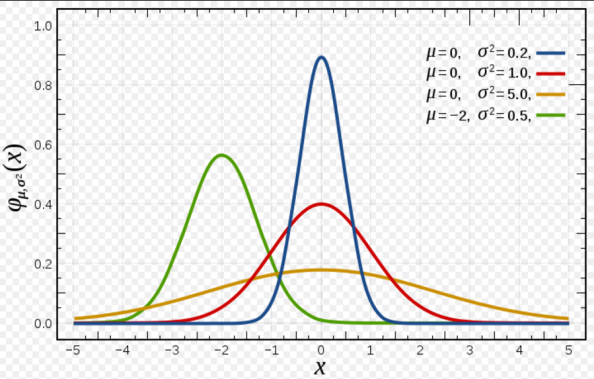
정규 분포에서의 확률은 변수의 평균 (μ) 및 표준 편차 (σ)에 따라 평균이 얼마나 벗어나는 지의 척도이다. 이 시점에서 변수와 평균 사이의 거리 인 Mahalanobis distance이다. 그리고 변수와 평균을 표준 편차로 나눈 값의 차이이다.

Figure 5. Normal Distribution (Univariate)

차원을 확장합시다. 이는 변수의 수를 늘리는 것을 의미한다. (단 변량 -> 다 변수)

다 변수에 대한 정규 분포의 확률 밀도는 다음과 같다.

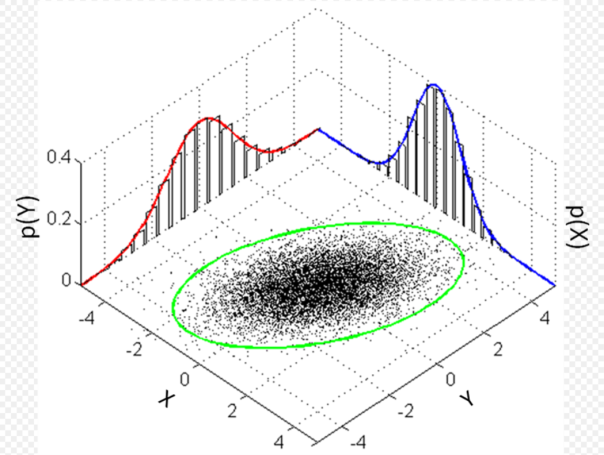
다차원에서도, 정규 분포에서의 확률은 변수가 평균으 로부터 어느 정도 벗어 났는지를 측정한다.

Figure 6. Normal Distribution (Multivariate)

데이터가 k 변수를 갖는 가우스 분포라고 가정하면, 데이터의 MD 제곱은 다음과 같이 k 자유도를 갖는 카이 제곱 분포를 따른다.

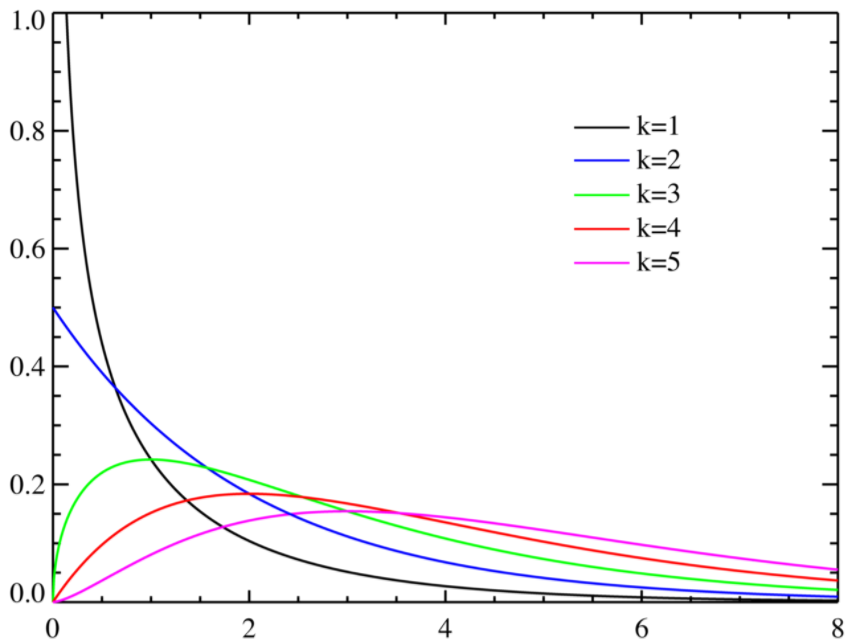
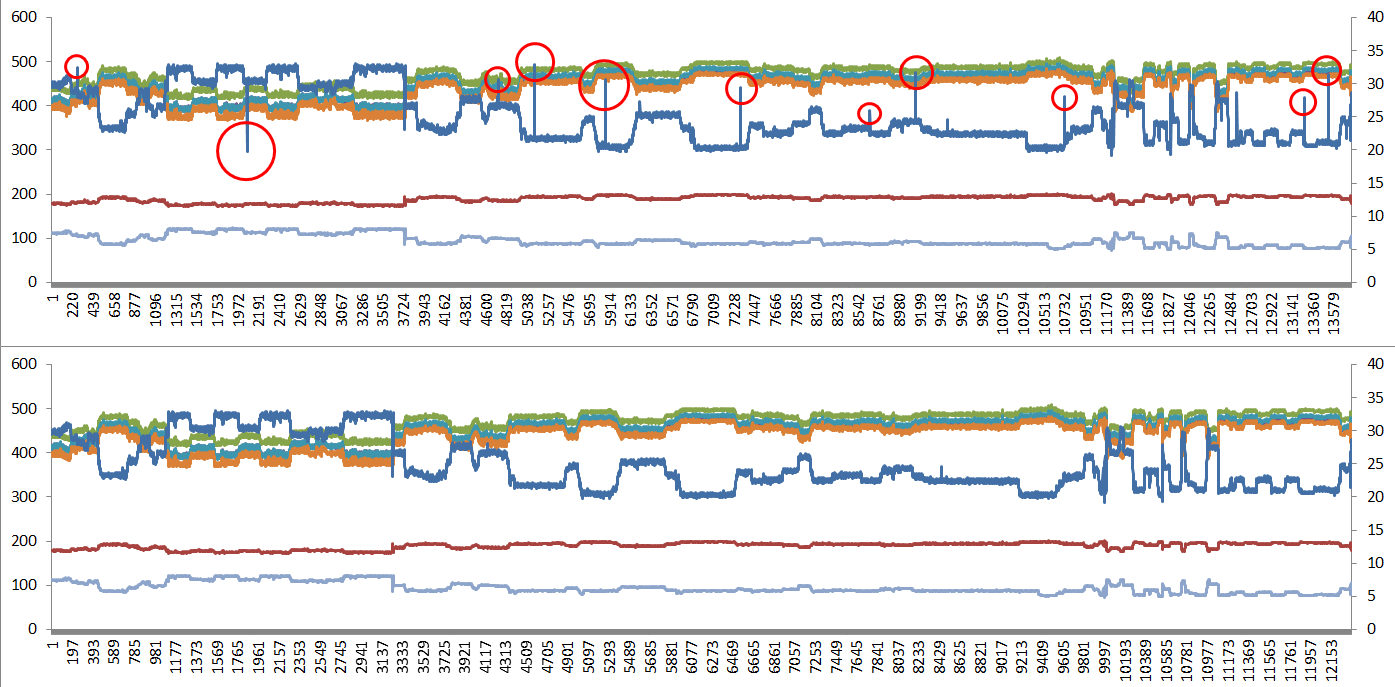


Figure 7. Chi-square Distribution

# Test Result

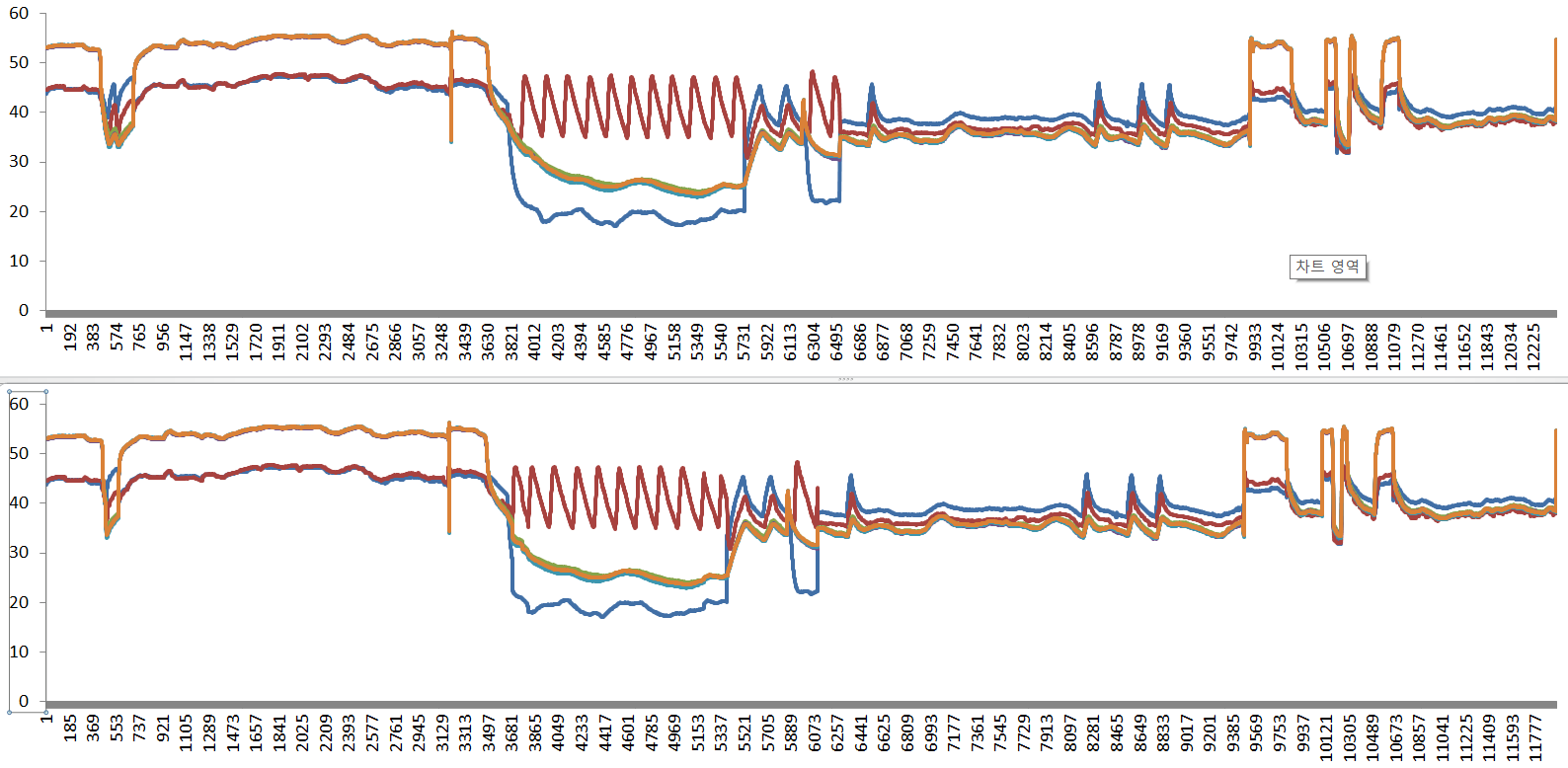
DJ03 모델 중에서 무작위로 10 개의 그룹을 선택하여 테스트하였다. 결과적으로 비정상적인 데이터가 제거되고 정상적인 데이터는 제거되지 않고 유지되었다.

**\* Test 1**



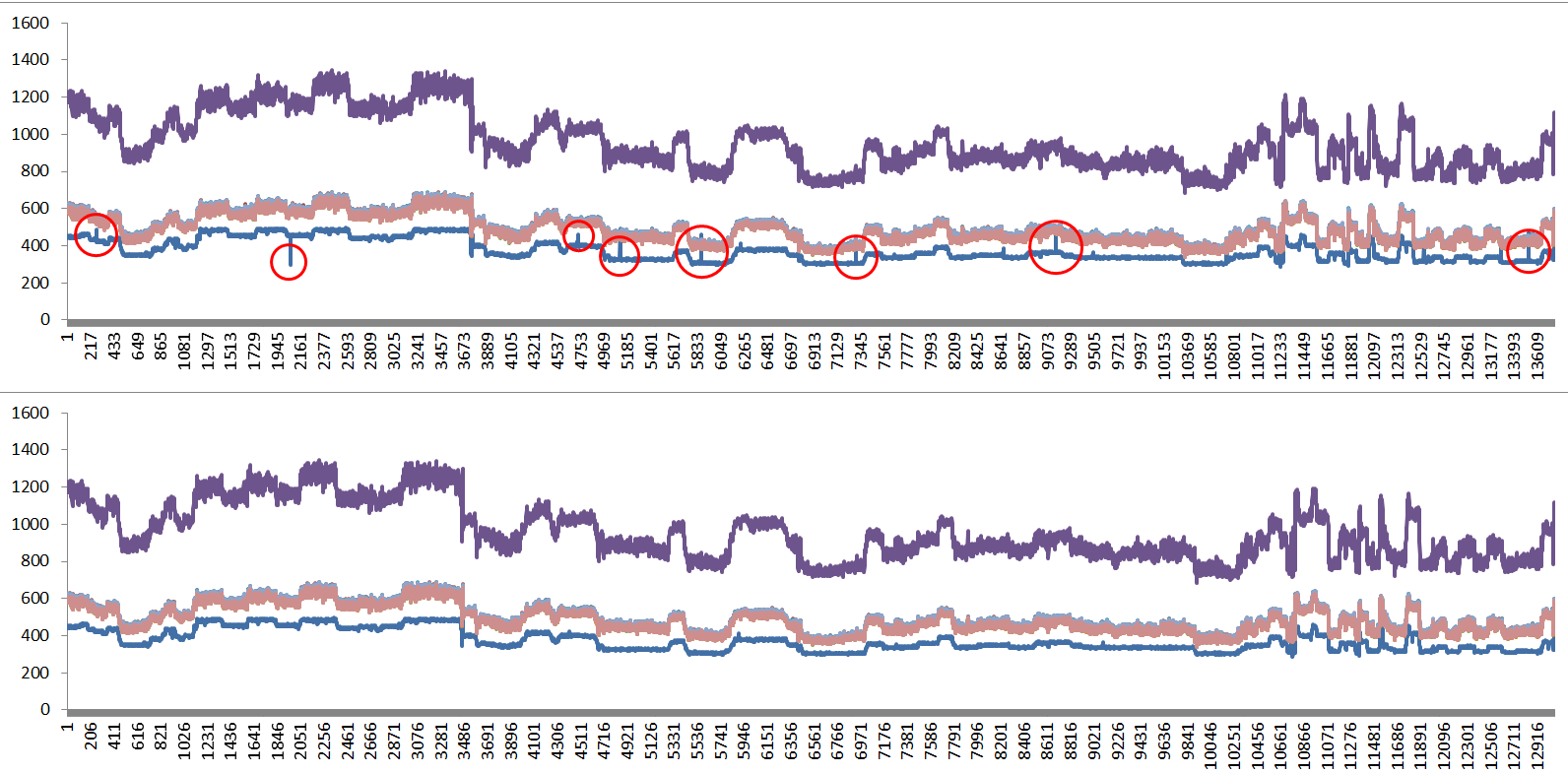
**After remove outliers**

**\* Test 2**



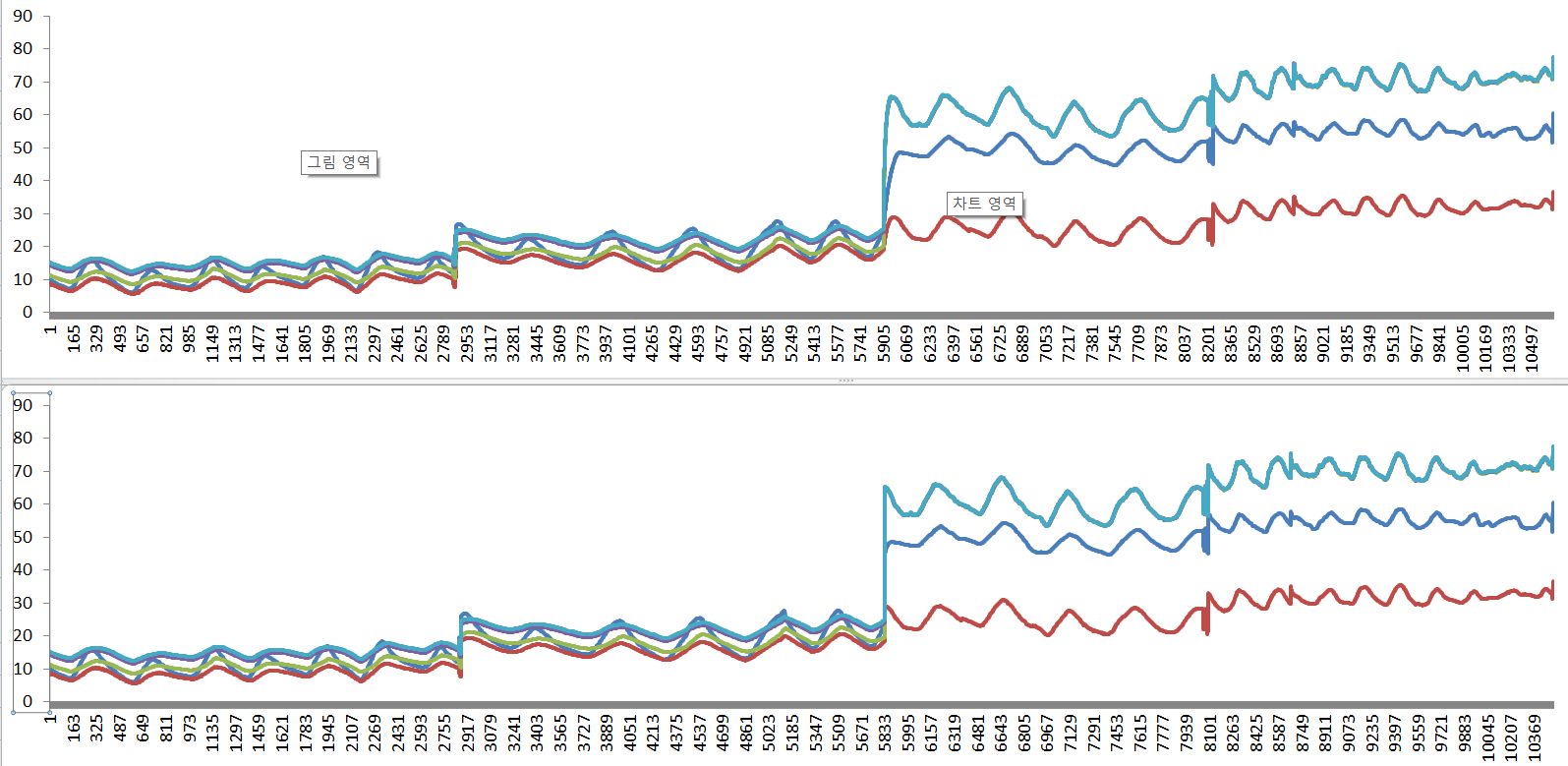
**After remove outliers**

**\* Test 3**



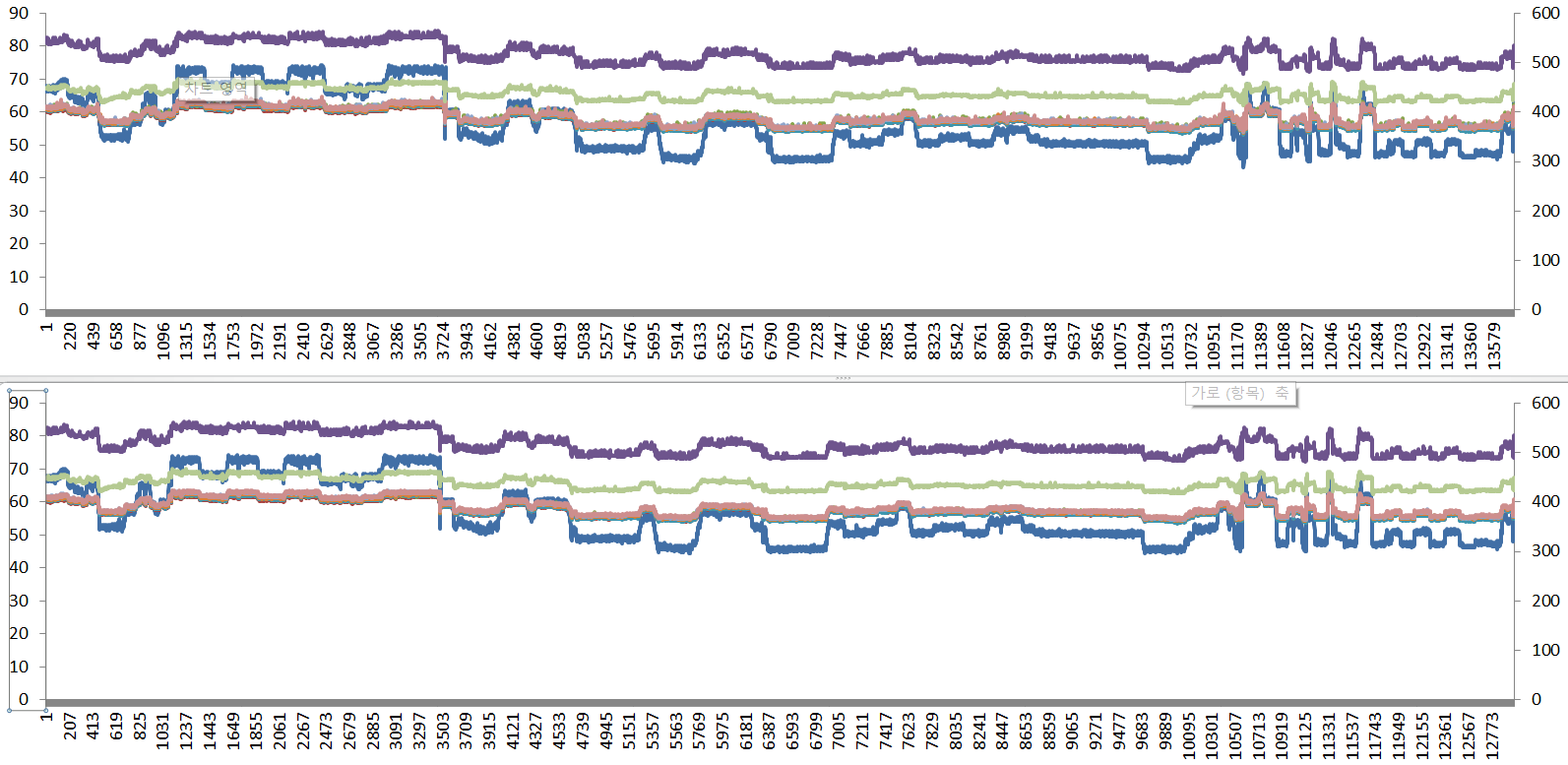
**After remove outliers**

**\* Test 4**



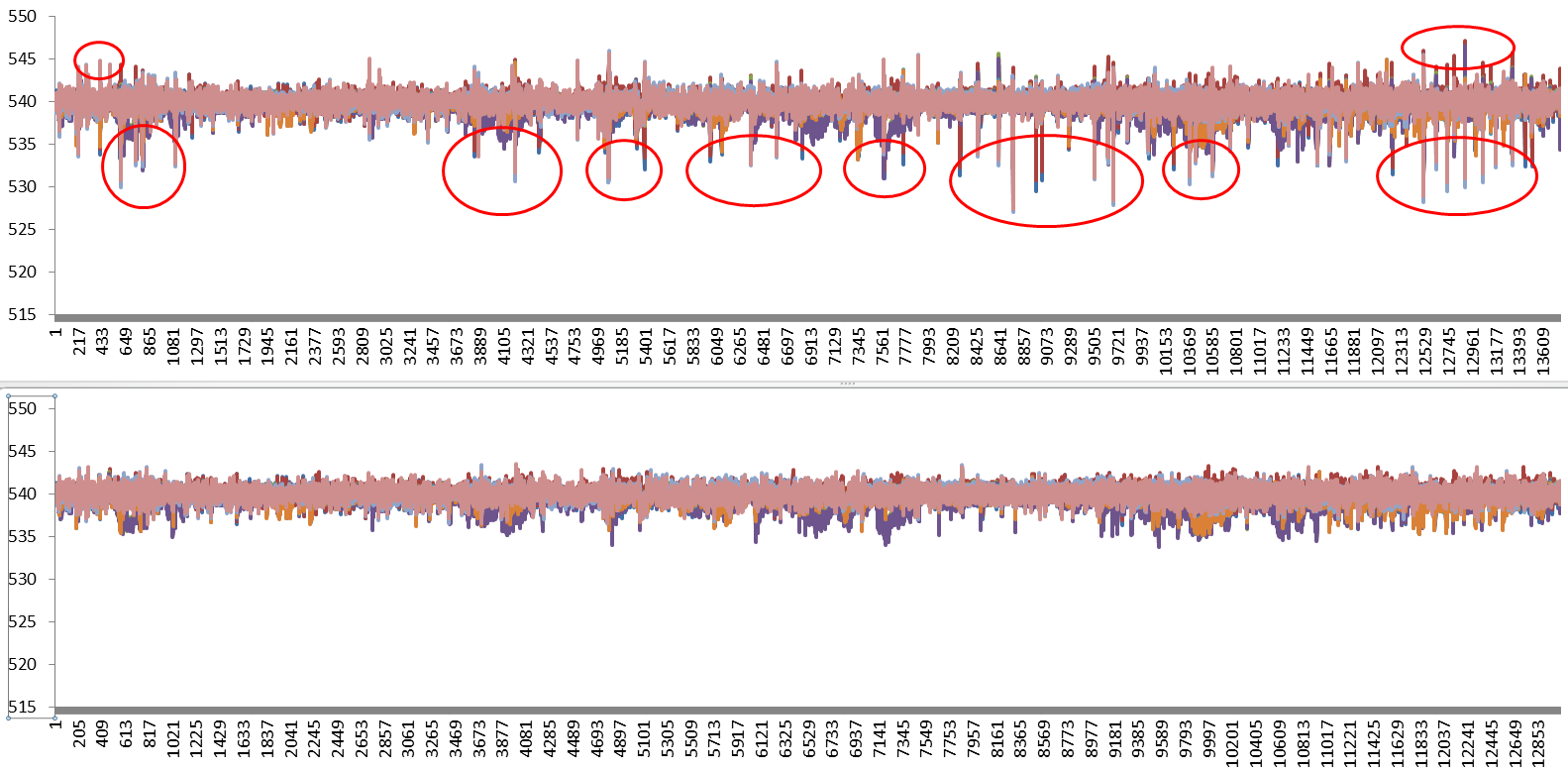
**After remove outliers**

**\* Test 5**



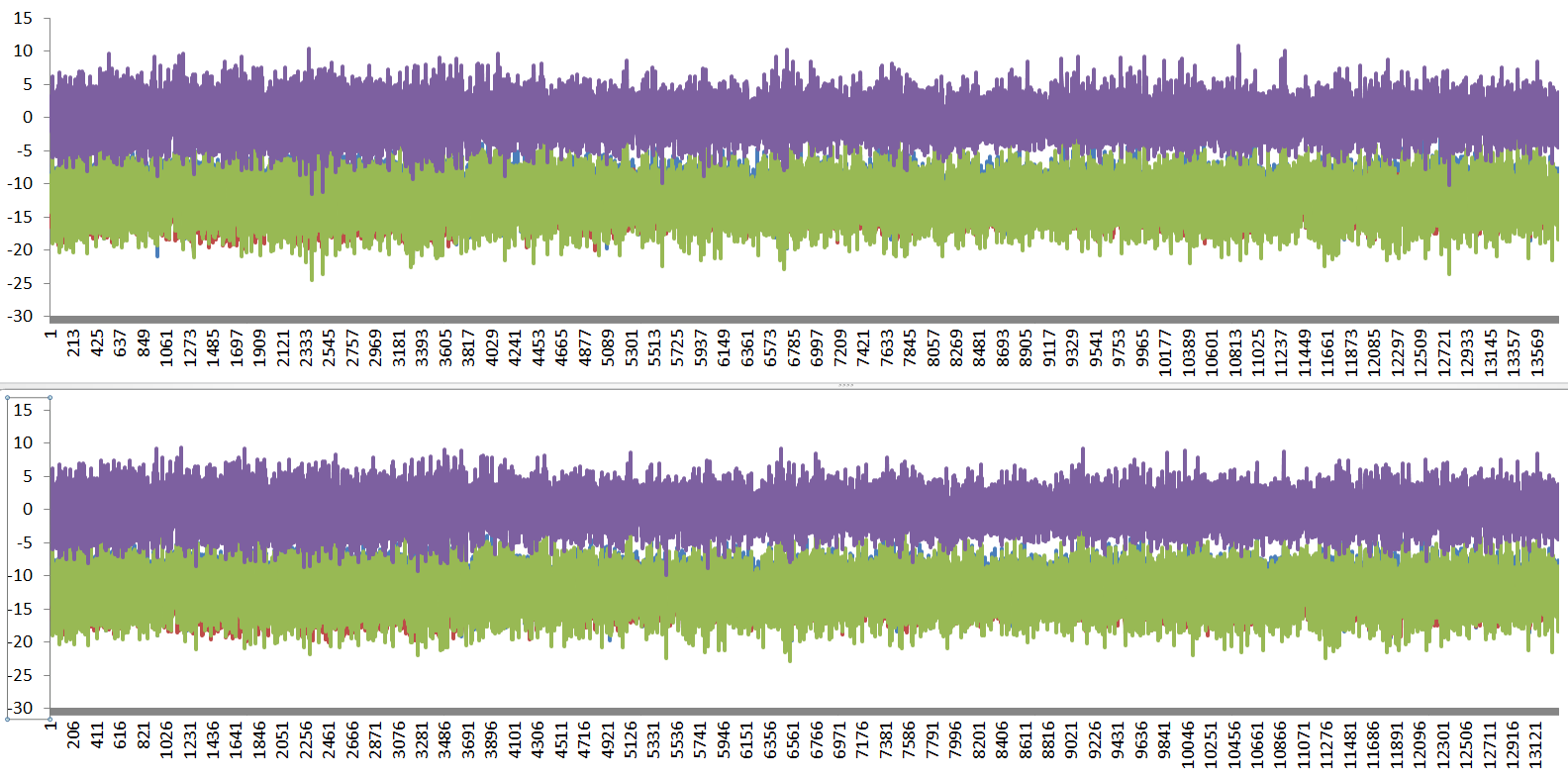
**After remove outliers**

**\* Test 6**



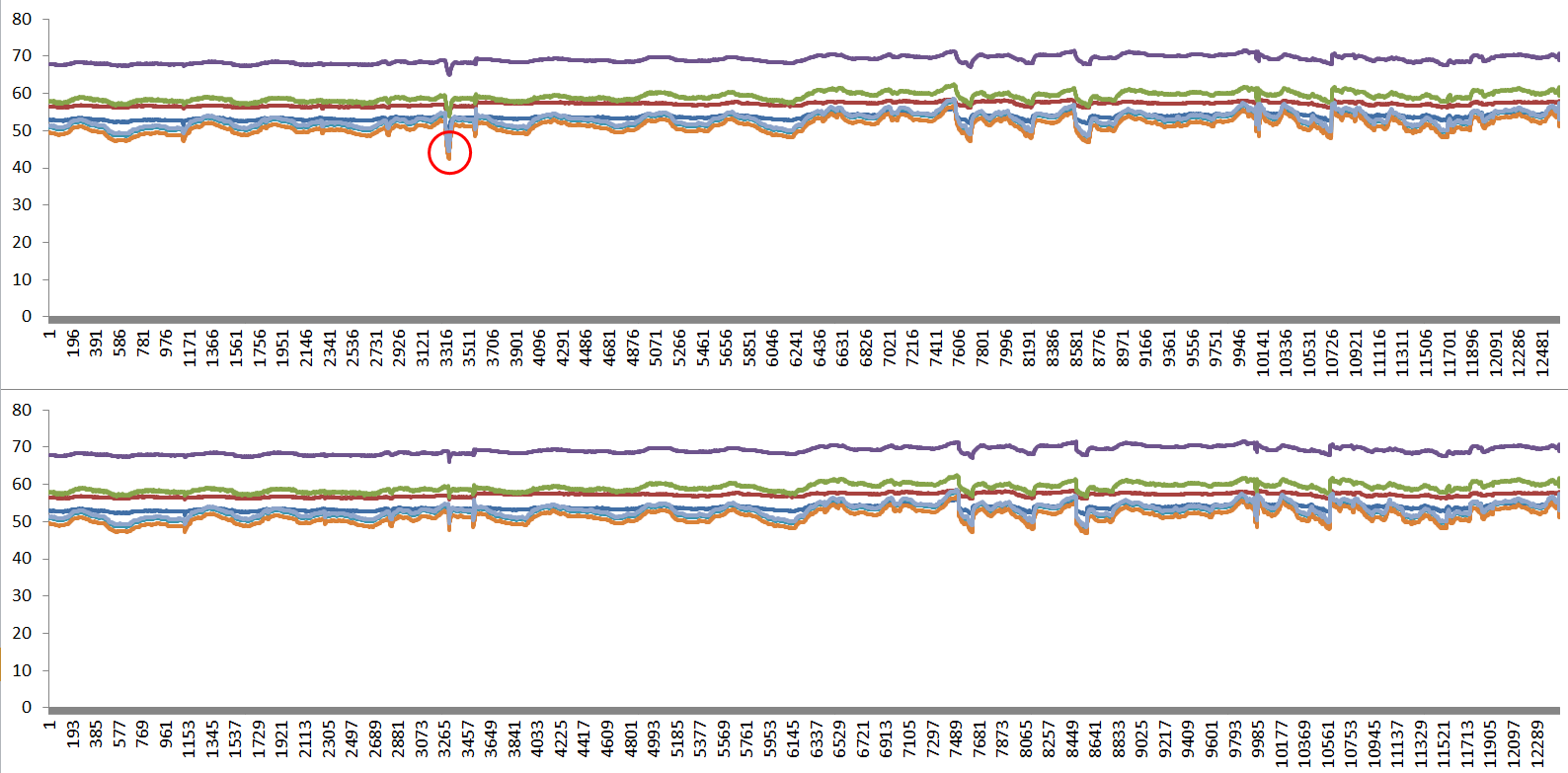
**After remove outliers**

**\* Test 7**



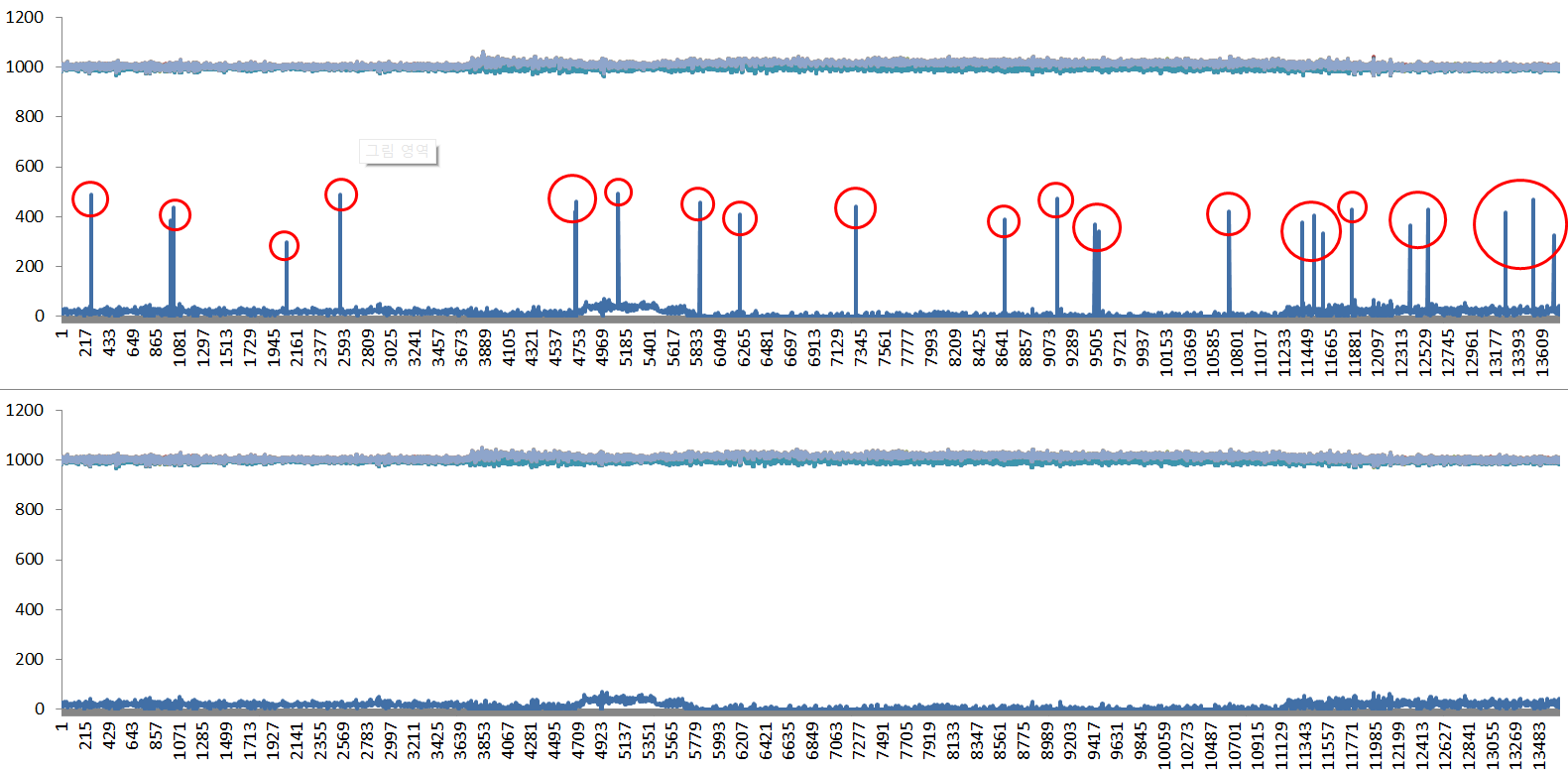
**After remove outliers**

**\* Test 8**



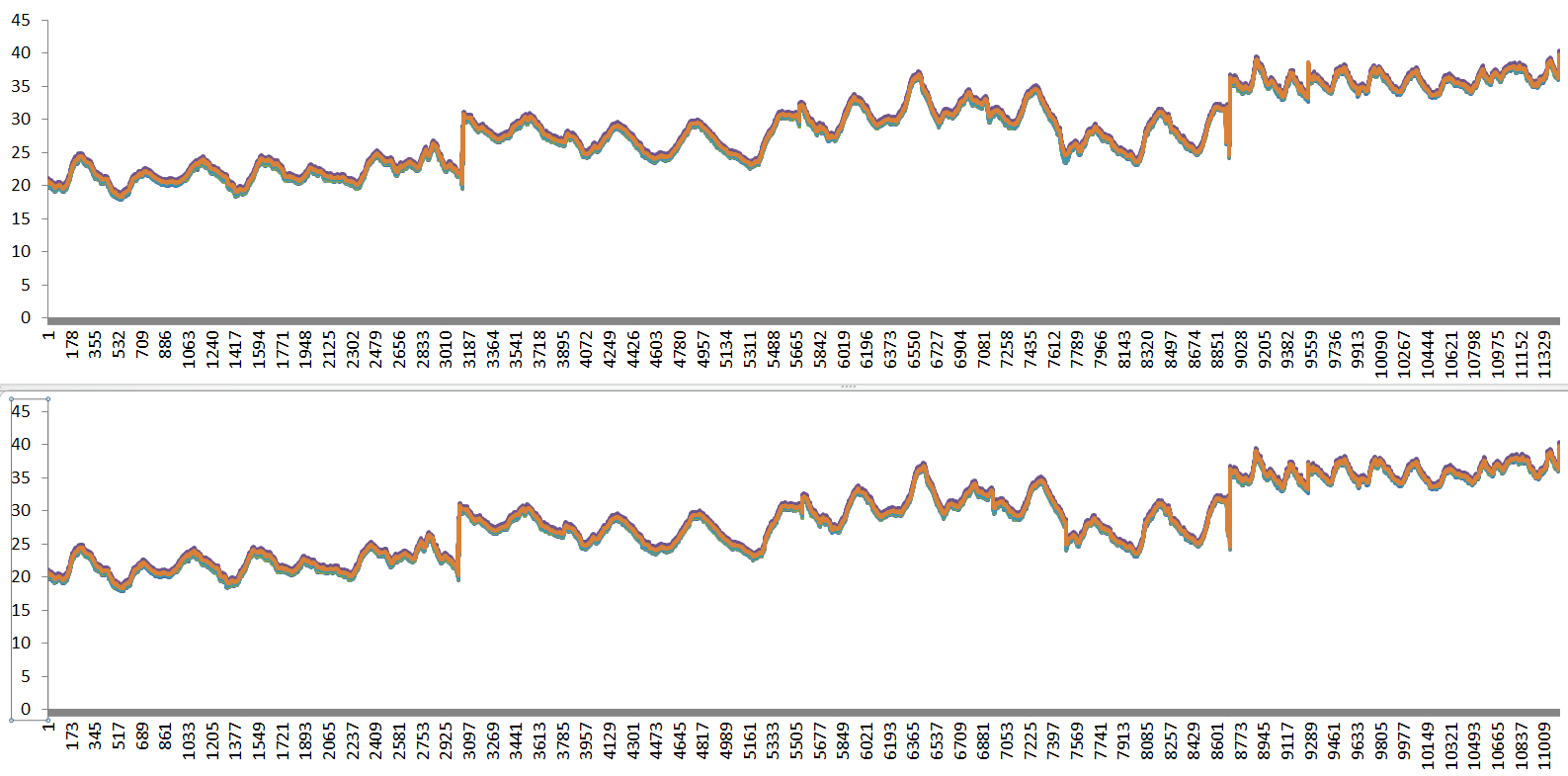
**After remove outliers**

**\* Test 9**



**After remove outliers**

**\* Test 10**



**After remove outliers**

# Expected benefits

비정상적인 데이터를 자동으로 제거함으로써 엔지니어링 시간을 효과적으로 단축하고 고객에게 잠재적 인 문제를 찾아 우리의 가치를 전달할 수 있다.