센서 데이터의 품질 개선을 위한 사용자 가이드

– 센서 데이터 정제 –

2025. 10

㈜지티원

목차

1 센서 데이터 품질 제고 필요성

1장은 크게 다음 두 내용을 소개하고 있다:

- 센서 데이터의 특성을 소개하고 이의 품질제고를 위한 노력이 왜 필요한가?
- 센서 데이터 품질 제고를 위해 개발된 ISO 표준(8000-210/220/230)은 어떤 구조로 어떤 내용을 담고 있는가?

1.1 센서 데이터의 특성 및 품질 제고 필요성

1.1.1 센서 데이터의 중요성

오늘날 센서 데이터는 다양한 산업과 현장에서 실시간으로 정확한 정보를 제공할 수 있다는 점에서 그 중요성이 날로 더해지고 있다. 센서 데이터의 중요성은 아래와 같이 구체화할 수 있다.

1) 업무 효율성 개선
   - 운용에 관한 실시간 정보를 지속적으로 제공해줄 수 있으므로 예방유지보수(preventive maintenance)을 가능하게 해준다. 이는 운영중인 시스템의 가동중단시간(downtime)을 줄임으로써 정비 비용을 감축할 수 있게 해준다.
   - 운용 과정에서 센서는 사용 여부와 사용 환경을 모니터링하여 에너지 소비를 최적화하여 상당한 비용 절감을 이룹니다.

2) 의사 결정 고도화
   - 센서는 다양한 산업에서 가치 있는 통찰력을 제공해줌으로써 이를 활용해 현명한 결정을 내리게 해줄 수 있다.

3) 안정성 제고
   - 센서는 업무 현장에서 안정 조건을 모니터링하는 데 중요한 역할을 수행한다.
   - 센서는 가스 누출, 화재, 구조적 불안정성과 같은 위험 상태를 감지해줌으로써 시의 적절한 대응과 사고의 사전 예방을 가능하게 해줄 수 있다.

4) 데이터 기반 혁신
   - 센서에 의해 수집된 대용량 데이터는 다양한 사업에서 혁신을 촉발시킬 수 있다. 이러한 혁신에는 신제품 개발, 새로운 서비스 창출, 또는 신규 사업 모델의 발현 등이 포함된다.

1.1.2 센서 데이터의 특성

센서 데이터는 전형적인 특징을 갖고 있다. 센서 데이터의 특징은 아래와 같이 구체화할 수 있다.

1) 실시간 또는 실시간에 가까운 성격
   - 센서는 모니터링하는 환경이나 시스템에 대한 최신 정보를 연달아 제공해준다.

2) 시점이 찍힌 정보
   - 센서 데이터의 각 측정점은 일반적으로 해당 시점의 정보가 달라붙어 있다. 이를 통해 센서 데이터는 시계열 (time series) 자료의 형태로 만들어진다.

3) 연속 흐름의 대용량 정보
   - 센서는 일반적으로 데이터의 연속적 흐름 정보를 만들어낸다. 즉, 시간이 지남에 따라 정보가 대용량으로 만들어진다.

4) 물리적 측정 정보
   - 센서 데이터는 온도, 압력, 움직임, 습도, 진동, 광선, 음향 등과 같은 물리적 양의 측정지표를 표현하고 있다.

5) 품질의 변동성
   - 원시 센서 데이터는 노이즈(noise), 불일치(inconsistencies), 데이터손실(data loss)과 같은 내용을 포함하는 경우가 많다. 그러므로 센서 데이터는 활용되기 전에 추가적인 처리가 요구된다.

1.1.3 센서 데이터와 일반 데이터의 차이점

센서 데이터는 우리가 이제껏 다루어 온 일반 데이터와는 여러 면에서 다른 특성을 갖고 있다. 이의 올바른 이해가 센서 데이터의 활용에 필수적이다. 센서 데이터와 기존 데이터와의 차이점을 아래와 같이 제시한다.

1) 출처
   - 센서 데이터는 물리적 환경으로부터의 입력을 식별하고 대응하는 장치에 의해 만들어진다.

2) 용량과 속도
   - 센서 장치는 대용량의 정보를 고속으로 만들어 낸다. 그러므로 이를 관리하고 분석하는 일 자체가 도전적인 업무에 속한다.

3) 구조적 포맷
   - 센서 데이터는 전형적으로 미리 정해진 속성으로 구성된 일관된 구조로 만들어지고 보관된다.

4) 사전 처리의 요건
   - 원시 센서 데이터는 일반적으로 추가적인 사전 처리가 요구된다. 여기에는 정제, 필터링, 정규화 등이 포함된다.

5) 상황 의존성
   - 센서 데이터의 값은 센서가 설치된 위치 또는 측정 시점의 주위 환경적 조건과 같은 상황에 종속된다.

1.1.4 센서 데이터 품질의 관리 및 제고 중요성

센서 데이터가 부정확하거나 신뢰할 수 없는 경우 의사결정의 결함, 시스템의 오작동, 심지어는 안전상 위험으로 이어질 수 있다. 그러므로 이런 센서 데이터 품질의 효과적 관리 및 품질 제고 노력이 중요하다. 그 구체적 이유는 다음과 같다:

1) 정확한 의사결정 수립

- 센서 데이터에 기초해 의사결정이 내려진다. 그런데 만일 데이터가 정확하지 않다면 그 의사결정은 결함이 있을 수밖에 없거나 값비싼 대가를 치르거나 위험한 상황에 직면할 수도 있다.
- 한 예로, 자율주행 차량에서 잘못된 센서 데이터는 부정확한 운행과 나아가서 충돌에 이를 수 있다.

2) 믿을 만한 시스템 수행
- 많은 시스템은 제대로 작동하기 위해 센서 데이터에 의존하는 경우가 많다. 품질이 낮은 센서 데이터 품질은 시스템 작동을 망가뜨릴 수 있다. 즉 오작동 또는 가동중단을 초래하기 쉽다
- 이런 문제는 시스템의 공정 제어를 어렵게 해 제품 하자 또는 비효율적 생산을 초래할 수 있다.

3) 안전 및 보안
- 센서 데이터는 자주 안전이 필수적인 분야에서도 사용된다. 위험상황을 감지하거나 인프라의 온전함을 모니터링 하는 경우가 여기에 속한다. 이런 면에서 부정확한 센서 데이터는 안전과 보안 관점에서 양보하는 경우가 발생할 수 있다.

4) 데이터 분석 및 통찰력
- 센서 데이터의 진정한 가치는 이 데이터의 분석을 통해 통찰력 정보를 제공하는 능력의 여부에 있다고 할 수 있다. 그러나 데이터가 오염되어 있거나 불완전한 경우 그 분석결과는 믿을 수 없게 되거나 통찰력 정보에도 아무 의미가 존재하지 않는다.

5) 효율성 제고 및 비용 절감
- 센서 데이터는 여러 분야의 운영 프로세스를 최적화하거나 효율성을 높이는 목적으로 자주 활용된다. 그러나 만일 데이터가 부정확하다면 최적화 노력은 그릇되게 되고 이는 자원의 낭비를 초래한다.
- 한 예로 스마트 빌딩의 센서가 공간 사용에 대한 잘못된 정보를 제공할 경우 냉난방환기 장치가 효율적으로 운영되지 못하는 문제가 발생한다.

1.2 센서 데이터 품질 제고를 위한 ISO 표준 개요

센서 데이터 품질 제고를 위한 ISO의 노력은 크게 세 가지 산출물로 구성된다.

- ISO 8000-210: 데이터 품질 — 210: 센서 데이터: 데이터 품질 특성
- ISO 8000-220: 데이터 품질 — 220: 센서 데이터: 품질 측정
- ISO 8000-230: 데이터 품질 — 230: 센서 데이터: 데이터 정제 가이드라인

ISO 8000-210 문서는 센서에서 생성된 데이터의 품질 특성(quality characteristics) 및 관련 데이터 이상 현상 (anomaly types)에 대해 설명한다. 그리고 ISO 8000-220 문서는 ISO 8000-210에서 규정된 품질 특성과 데이터 이상을 정량적으로 측정하는 품질 지표(quality measure)를 정의한다. 이 두 문서의 내용을 기초로 하여 ISO 8000-230 문서는 데이터 품질을 개선하기 위해 데이터 품질에 영향을 미치는 데이터 이상(anomaly)을 정제 (data cleansing)하는 지침을 다룬다. 이 지침은 센서 데이터 정제 과정의 원칙, 절차 및 구현 요구사항을 포함한다. 이 과정은 ISO 8000-210에서 정의된 데이터 품질 특성 및 이상과 ISO 8000-220에서 정의된 데이터 품질 측정 지표를 기반으로 센서 데이터 정제를 수행한다. (이 문서들에서 사용되는 용어 정의는 부록 A를 참조)

이들 간의 관계는 그림 1-1과 같이 나타낼 수 있다.

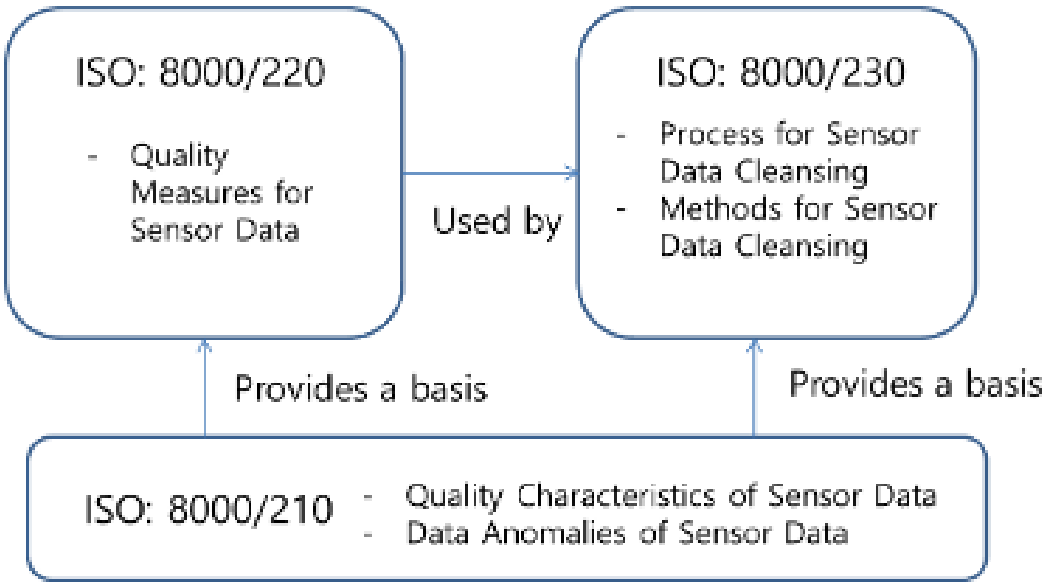


그림 1: ISO 8000-210/220/230 간의 관계

1.2.1 센서데이터 품질 특성 및 측정의 개념적 요소에 대한 이해

센서데이터의 품질 특성, 이상 및 측정에 대한 개념 요소 간의 관계가 그림 1-2에 묘사되어 있다.

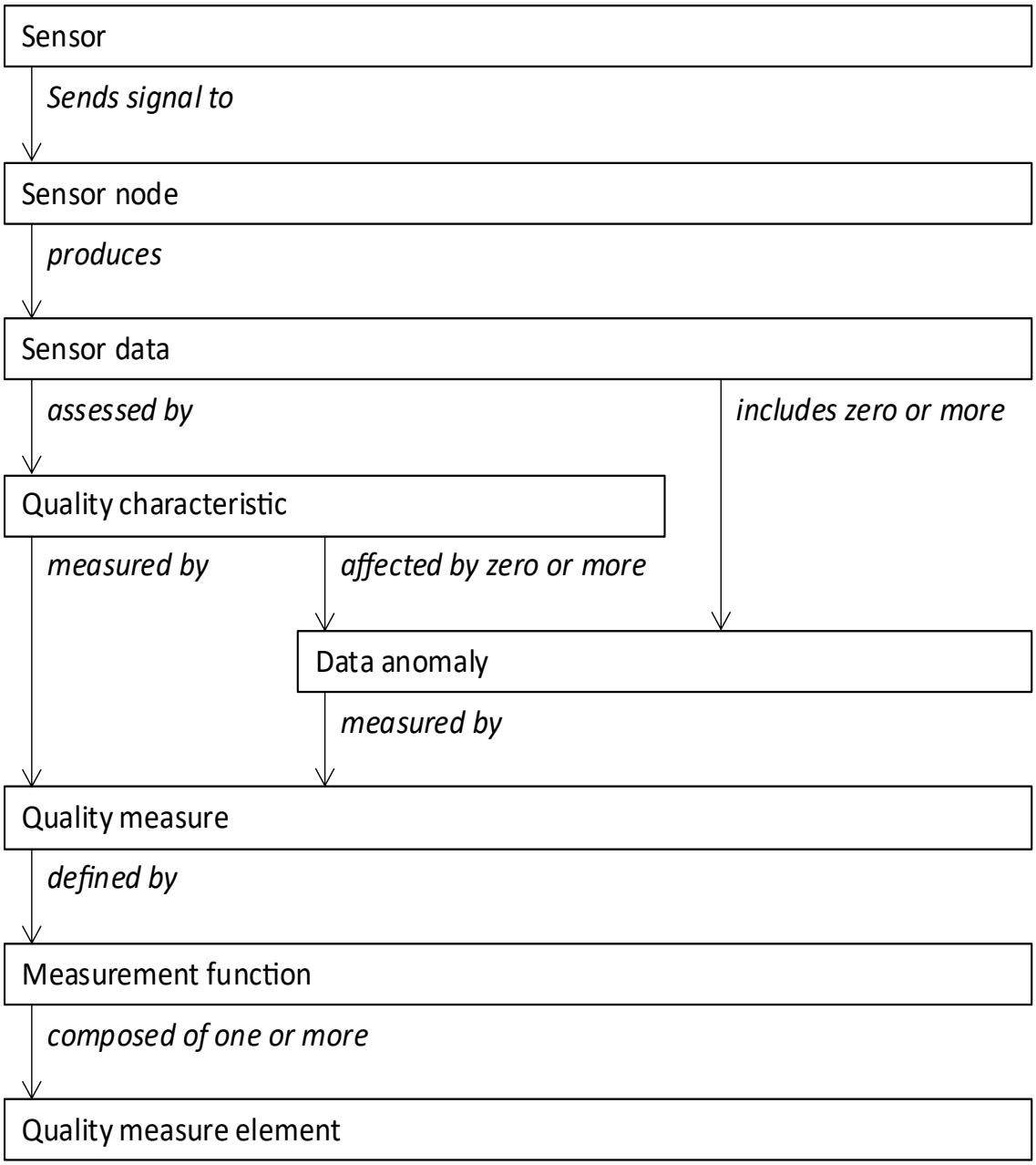


그림 2: 센서데이터 품질의 개념 요소 및 관계

1.2.2 센서 데이터 품질 제고를 위한 ISO 표준 적용 분야

센서 데이터 품질 제고를 위한 ISO 표준은 이산형 스트림 데이터 (discrete stream data) 형태로 제공하는 센서가 설치된 모든 시스템 또는 장비에 적용 가능하다.

그러나 다음과 같은 형태의 센서에는 적용되지 않는다:

- 이미지, 오디오, 비디오 등의 미디어 형태로 제공되는 센서
- 아날로그 형태로 제공되는 센서
- 아날로그 데이터를 디지털 데이터로 생성하기 위해 변환하거나 수정하는 신호 처리;

센서 데이터 품질 제고를 위한 표준 및 방법론은 다양한 산업에서 다양한 목적으로 활용될 수 있다고 본다. 아래 제시한 분야가 대표적으로 활용될 가능성이 높다. 여기 언급되지 않은 산업에서도 향후 다양한 분야에서 적용될 수 있으리라 예상한다.

1) 제조업
   - 제조 공정에서 센서 데이터 품질 향상은 불량률 감소와 공정 안정성 증대로 직결되어 생산성과 신뢰성 모두 큰 개선 효과를 가져온다.
   - 공정 세분화, 생산 품질 관리 효율화, 설비 예지 보수 등에서 데이터 정확성가 중대하게 작용한다.
2) 스마트시티/인프라
   - 스마트시티에서 IoT 센서 데이터를 통한 교통, 환경, 에너지 관리의 효율성 및 공공 서비스 품질이 크게 향상된다.
   - 고품질 데이터는 도시 인프라의 실시간 모니터링 · 의사결정, 시민 안전 확보 등에도 중요하다.
3) 수자원 관리/환경 모니터링
   - 수질, 수량, 기상 등 환경 관련 분야는 센서 데이터가 정책 결정과 위기 대응에 직접 연결되어, 데이터 품질 향상 시 예측 모형과 의사결정이 크게 고도화된다.
4) 스마트 헬스케어
   - 센서 데이터 품질이 개선되면 웨어러블, 스마트워치, 원격 모니터링 기기에서 얻는 생체신호(심박, 혈압 등) 기반 진단과 건강 모니터링 정확성이 향상되어 오진이나 오경보, 데이터 누락으로 인한 위험을 줄일 수 있다.
   - 의료 데이터 오류 최소화는 환자 안전 및 신속한 의사결정에 직접적 영향을 미친다.
   - 적용 시스템 환경(수정 요)

1.2.3 센서 데이터 품질 제고를 위한 ISO 표준 적용 시스템 환경

센서 데이터는 센서 네트워크 또는 사물 인터넷(IoT)에 연결된 하나 이상의 센서에서 생성된다. 이러한 네트워크(사물 인터넷 포함)는 데이터 수집, 변환, 전송, 저장 및 분석을 포함한 일련의 데이터 관련 기능을 포함한다. 이러한 기능은 데이터 품질이 적절한 수준일 때 더욱 효과적이다.

일반적으로 센서 네트워크는 크게 세 가지 영역으로 구조화되어 있다고 볼 수 있다. 그림 1-3에서 보는 바와 같이, 신호 데이터가 생성되는 “Sensing domain”, 해당 데이터를 필요한 시스템, 조직 등으로 전송되는 “Network domain”, 그리고 목적에 따라 가공/분석되어 활용되는 “Service domain” 으로 구

조화된다. 이 세 가지 영역 중에서 본 표준은 "Sensing domain"과 "Service domain"에서 활용될 수 있다.

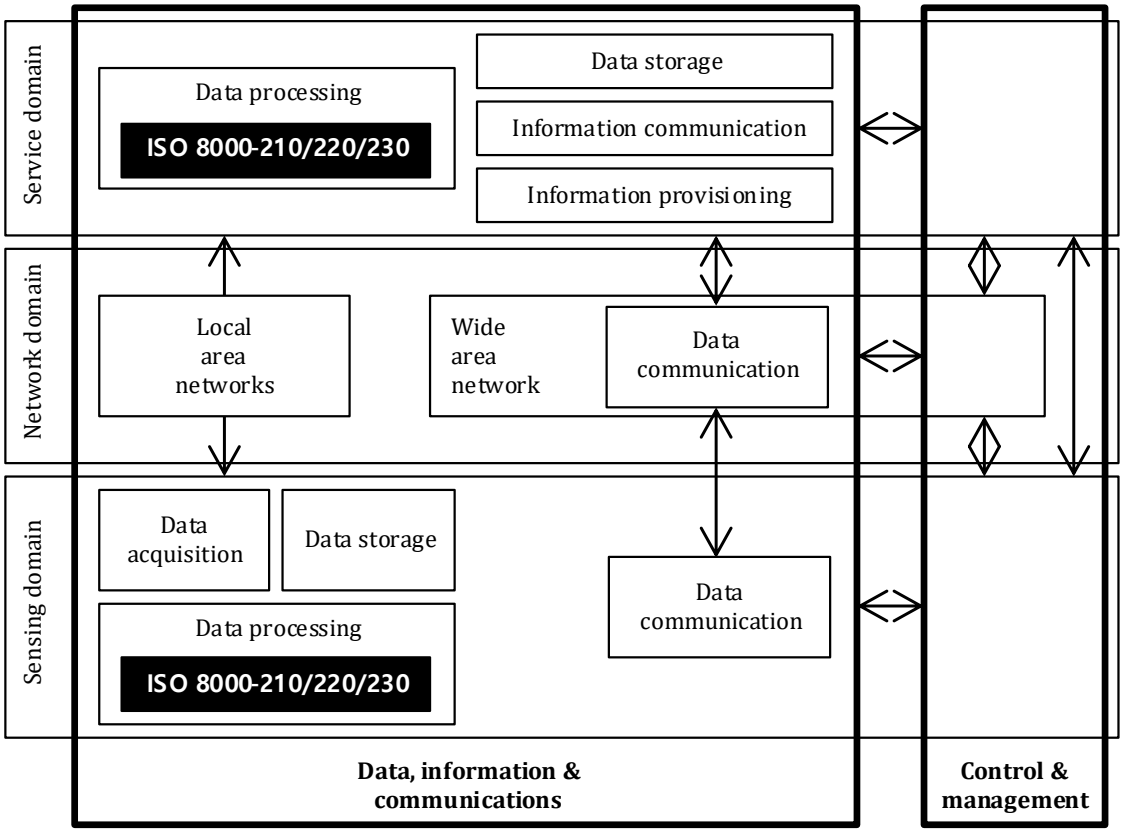


그림 3: 센서 네트워크 환경에서의 ISO 8000-210/220/230의 역할

2 센서 데이터의 이상 유형

본 장에서는 ISO 8000-210에서 정의하고 있는 센서 데이터의 이상 유형을 최대한 자세히 기술한다. 센서 데이터는 물리 · 환경적 요인, 센서 자체의 성능 한계, 네트워크 · 시스템 장애 등에 의해 종종 정상 범주를 벗어난 패턴을 보이는데, 이를 이상(anomaly)이라 하며, 이러한 이상을 식별 및 관리하는 것은 데이터 품질(정확성, 완전성, 일관성, 정밀성, 적시성)을 유지 및 제고하는 데 매우 중요하다.

ISO 8000-210은 크게 개별 센서의 이상과 복수(복합) 센서 간의 이상으로 구분하고 있으며, 각 유형 및 정의는 다음과 같다.

2.1 개별 센서의 이상

개별 센서 수준에서 발생하는 이상은 센서 한 대만 놓고 보았을 때 관찰할 수 있는 비정상적 패턴이나 값의 형태를 의미한다. ISO 8000-210의 5.2에서는 대표 유형들을 다음과 같이 제시한다.

2.1.1 오프셋

오프셋(offset)은 수집된 데이터 값이 기대 값과 항상 일정한 간격으로 벗어나는 이상을 의미한다.

- 예: 실제 온도가 25°C일 때, 센서는 항상 27°C로 기록하는 경우.

2.1.2 드리프트

드리프트(drift)는 센서 데이터 값이 시간이 지남에 따라 연속적으로 증가하거나 감소하는 등, 실제 값과의 편차가 누적되는 현상이다.

- 초기에는 정상 범위에 있으나 시간이 지날수록 오차가 점차 커짐.
- 예: 초기에는 -0.5°C 오차였는데, 1시간 후에는 -2°C 오차가 되는 식으로 점진적 편차가 발생.

2.1.3 트림

트림(trim)은 센서가 일정 범위를 넘어선 실제 값을 제대로 표시하지 못하고, 상한 또는 하한 값으로 '잘려서(trimmed)' 기록되는 현상이다.

- 센서 측정 임계값(threshold) 이상(또는 이하)의 실제 값이 동일한 최대(또는 최소)값으로만 저장됨.
- 예: 0~1000 범위까지만 측정 가능한 센서에서 1200을 측정해야 할 상황이 발생하면, 실제 데이터는 1000으로 '잘려서' 기록됨.

2.1.4 스파이크

스파이크(spike)는 정상 범주로부터 단발성 혹은 일시적으로 급격하게 튀는(돌출되는) 값(outlier)이 나타나는 현상이다.

- 한두 지점에서만 극단적 수치가 발생한 후, 많은 경우 다시 정상 범위로 복귀.
- 예: 노이즈, 통신 에러, 일시적 간섭 등으로 인해 단 한 번 크게 튀는 값이 포착되는 경우.

2.1.5 노이즈

노이즈(noise)는 센서 데이터가 무작위 잡음으로 인해 정상 값 주변에서 과도하게 흔들리거나 균질하지 못한 분포를 보이는 현상이다.

- 센서 하드웨어 성능, 주변 환경 간섭, 무선 신호 간섭 등으로 발생. 값이 지속적으로 미세하게 흔들리며 분석 시 정확성이 떨어질 수 있음.
- 예: 센서가 냉기 배수 시스템에 대한 데이터를 수집하는데 하드웨어 오류로 인해 데이터의 변동폭이 예상보다 크다.

2.1.6 데이터 손실

데이터 손실(data loss)은 센서가 데이터를 전송 기록해야 할 시점에 누락(missing)되어 측정값이 비어있거나 기본값(예: 0)만 기록되는 현상이다.

- 네트워크 단절, 배터리 부족, 센서 재부팅 등의 원인으로 발생.
- 예: 1초 간격으로 10개의 데이터를 받아야 하는데, 실제로 2개의 데이터가 누락된 경우.

2.1.7 데이터 양 부족

데이터 양 부족(lack of amount)은 특정 시간 동안, 분석이나 활용 목적을 만족하기에 충분한 수량의 데이터가 확보되지 못하는 현상이다.

- 실질적으로 데이터 손실과 비슷하지만, 데이터 수집 빈도 및 횟수가 요구 수준에 못 미치는 상태를 의미.
- 예: 진동 예측 모델은 초당 100회의 측정값이 필요한데, 센서 사양상 초당 50회만 측정 가능한 경우.

2.1.8 편이

편이(shift)는 센서 데이터가 일정 추세를 유지하다가 갑자기 다른 추세(또는 다른 평균값)로 변하여 그 패턴을 벗어나는 현상이다.

- 오프셋이 시간 축 전체에 걸쳐 일정하게 발생하는 반면, 편이는 특정 시점에 불연속적으로 발생.
- 예: 10분간 평균 50 근처를 유지하던 측정값이 순간적으로 70 근처로 올라가더니 계속 그 주변에서 유지되는 경우.

2.1.9 급락/급상승

급락/급상승(drop or rise)은 센서 데이터가 단기간에 비정상적으로 큰 폭으로 하강 혹은 상승하는 현상이다.

- 짧은 구간에 걸쳐 가파르게 변동하며, 이후 다시 정상범위로 복귀하거나 다른 상태로 전이.
- 예: 센서 배터리 상태나 기기 오류로 인해 순간적으로 0 근처(급락) 또는 매우 큰 값(급상승)으로 표출되는 상황.

2.1.10 값 고정

값 고정(stuck)은 센서가 특정 시간 동안 거의 변동 없이 동일 값 혹은 근접 값만 계속 기록하는 현상이다.

- 실제로는 시간이 지날수록 값이 변해야 하는 상황임에도, 센서 오류나 블로킹 등으로 인해 측정값이 멈춰 있음.
- 예: 외부 충격으로 센서 측정부가 고장 나서 계속 같은 수치만 전송.

2.1.11 범위 진동

범위 진동(bound oscillation)은 센서가 기록하는 값의 전체 범위(range)가 특정 시점 이후 갑자기 달라지거나 바운더리가 변화하는 현상이다.

- 센서 설치 위치 변경, 교정(calibration) 과정에서 범위 세팅이 바뀌는 경우 발생.
- 예: 13 범위를 측정하던 센서가 어딘가 이동된 후 48 범위에서 측정값이 생성됨.

2.1.12 주기 불일치

주기 불일치(inconsistent frequency)는 원래 설정된 측정 주기가 유지되지 않고, 중간에 변동되거나 고르지 못한 간격으로 데이터가 생성되는 현상이다.

- 구성(설정) 변경, 센서 노드 소프트웨어 업그레이드 등으로 샘플링 레이트(rate)가 달라질 수 있음.
- 예: 10Hz로 측정 중이던 센서가 어느 순간부터 1Hz, 혹은 50Hz로 바뀌어 데이터를 수집하는 경우.

2.1.13 해상도 불일치

해상도 불일치(different resolution)는 센서가 시간대나 상태에 따라 서로 다른 소수점 자릿수 혹은 분해능으로 데이터를 기록하는 현상이다.

- 특정 값 범위를 넘어가면 소수점 셋째 자리까지만 표시한다거나, 펌웨어 업데이트 후 해상도가 달라질 수 있음.
- 예: 0~99 구간에서는 xxx.xx(소수 둘째 자리)까지, 100 이상부터는 xxx.x(소수 첫째 자리)까지만 기록.

2.1.14 잘못된 시간정보

잘못된 시간정보(incorrect timestamp)는 센서 데이터에 기록된 시간 정보가 실제 측정 시점과 불일치하거나 비합리적인 값을 갖는 현상이다

- 센서 노드의 RTC(Real-Time Clock) 오류, 통신 지연, 시간 동기화 실패 등으로 발생.
- 예: "2224-12-01 12:30"처럼, 현재 시점과 전혀 관계없는 미래 시간이 찍히는 경우.

2.1.15 지연

지연(latency)은 센서가 데이터를 발생(또는 전송)해야 할 시점보다 훨씬 늦게 데이터를 전달하거나 저장하여, 시계열의 적시성을 해치는 현상이다.

- 고사양 연산 요구, 네트워크 지연, 버퍼링 문제 등으로 인해 생길 수 있음.
- 예: 10ms 내에 전송돼야 할 데이터가 1~2초 늦게 들어오는 경우, 실시간 제어에 치명적 문제.

2.2 복합(다중) 센서의 이상

센서 여러 대가 동시에 혹은 상호 연관된 방식으로 데이터를 측정·수집할 때, 센서 간 데이터가 불일치하거나 불규칙하게 나타나는 현상을 말한다. ISO 8000-210의 5.3에서는 대표적으로 다음 유형을 언급한다.

2.2.1 이질성

이질성(dissimilarity)은 동일 대상(또는 물리량)을 측정하는 복수 센서가 '원래 같은 값을 보여야 하는' 맥락임에도, 상당히 다른 수치를 기록하여 상호 불일치가 두드러지는 현상이다.

- 센서 간 단위(unit) 또는 교정(calibration) 차이, 하드웨어 오차 편차 등으로 인해 발생.
- 예: A 센서와 B 센서가 동시에 온도를 측정했는데, A는 25.0°C, B는 30.5°C로 기록.

2.2.2 규칙 위반

규칙 위반(Rule violation)은 서로 연관된 센서 간에 사전에 정의된 관계(규칙, 수식 등)를 만족해야 함에도, 이를 위배하는 데이터가 관찰되는 현상이다.

- 예 1: A 센서 + B 센서 = C 센서라는 공정상 불변 관계가 있는데, 실제 측정값이 A+B ≠ C를 지속적으로 나타내는 경우.
- 예 2: 전력 계산, 유량 및 압력 균형, 센서 융합(fusion) 등에 흔히 발생.

2.2.3 시간정보 불일치

시간정보 불일치(inconsistent timestamp)는 복수 센서가 동시에 같은 시점의 물리량을 측정해야 함에도, 기록된 시간 정보가 서로 달라 동기화가 깨진 듯 보이는 현상이다.

- 하나의 이벤트(예: 기계동작 시작)를 측정해야 하는데, 센서 A에는 12:00:00.500, 센서 B에는 12:00:02.000 등으로 차이가 큼.
- 예: 네트워크 지연, 센서 노드 간 시간 동기화 알고리즘 불일치가 원인.

2.3 이상 유형이 센서 데이터 품질 특성에 미치는 영향

앞에서 열거한 이상 유형들은 3장에서 정의된 센서 데이터 품질 특성(정확성, 완전성, 일관성, 정밀성, 적시성)에 직접적으로 영향을 준다. 예컨대, 오프셋, 드리프트, 스파이크는 주로 정확성 저하를 야기하며, 데이터 손실이나 데이터 양 부족은 완전성에 문제를 일으킨다. 센서 간 규칙 위반이나 이질성은 데이터 일관성에 직접적인 영향을 준다. 이처럼 각 데이터 이상은 복수의 품질 특성에 복합적 영향을 줄 수도 있고, 반대로 여러 가지의 데이터 이상들이 특정 품질 특성에 심각한 영향을 줄 수도 있다.

3 센서 데이터의 품질 특성

3.1 개요

센서 데이터는 시간에 따라 연속적으로 측정되고, 외부 환경 · 장비 상태 · 통신 상황 등에 의해 다양한 이상을 겪을 수 있다. 이러한 센서 데이터가 실제 활용 목적(분석, 제어, 예측 등)에 적합한지를 평가하기 위해서는, 품질 특성(quality characteristics)을 정의하고 이에 대한 충족도를 측정하는 과정이 필수적이다.

ISO 8000-210 에서는 센서 데이터에서 중요하게 다뤄야 할 품질 특성으로 다음 다섯 가지를 제시한다.

- 정확성(accuracy)
- 완전성(completeness)
- 일관성(consistency)
- 정밀성(precision)
- 적시성(timeliness)

이러한 품질 특성은 ISO 8000-220에서 정의한 품질 측정 지표(quality measures)를 통해 정량적으로 평가할 수 있으며, ISO 8000-230에서 제시하는 데이터 정제(data cleansing) 가이드와 결합하여 데이터의 품질을 개선하는 근거로 활용된다.

ISO 8000-210 에 따르면, 센서 데이터의 품질을 평가하기 위해 고려해야 할 대표적 특성은 정확성, 완전성, 일관성, 정밀성, 적시성이며, 이들은 서로 맞물려 센서 데이터 전반의 신뢰도를 결정한다. 또한 특정 애플리케이션에서 요구하는 기준과 목적에 따라, 다섯 가지 특성 중 어느 하나에 더 높은 가중치를 둘 수도 있다.

예를 들어, 실시간 제어가 중요한 산업 현장에서는 적시성이 가장 우선시될 수 있고, 의료용 진단 장치에서는 정확성과 정밀성이 절대적으로 중요할 수 있다.

3.2 정확성

1) 정의

정확성(accuracy)은 센서 데이터가 의도된 개념 혹은 이벤트의 참값(true value)을 얼마나 올바르게 반영하고 있는가를 나타내는 정도

- 예: 실제 온도 25.0°C를 측정해야 하는 상황에서 센서가 25.1°C 혹은 24.9°C를 기록한다면, 이는 높은 정확성으로 볼 수 있으나, 28°C 정도로 기록한다면 정확성이 낮다고 할 수 있다.

2) 관련 이상

오프셋, 드리프트, 트림, 스파이크, 노이즈 등이 정확성에 큰 영향을 미친다.

- 오프셋: 모든 측정값이 특정 상수값만큼 참값과 차이 나는 현상
- 드리프트: 시간이 지남에 따라 오차가 커지거나 누적되는 현상
- 스파이크: 일시적으로 급격하게 튀는 값이 발생하여 평균을 왜곡

- 노이즈: 랜덤 잡음으로 인해 측정값이 참값 주위에서 불규칙하게 흔들리는 경우 등

3) 품질 측정

데이터 세트의 정확성 = A / B

A = 특정 허용 오차 범위 안에 들어오는 데이터 값의 개수
B = 전체 데이터 개수

따라서 A/B가 높을수록 정확성이 높다는 의미이다.

3.3 완전성

1) 정의

완전성(completeness)은 센서 데이터가 특정 맥락(시간적·구간적 요구사항)에서 필요한 모든 값을 누락 없이 확보했는가를 나타내는 정도.

- 예: 1초 간격으로 100개의 데이터가 필요하다고 했을 때, 실제로 95개만 수집되었다면 완전성은 95%에 해당한다.

2) 관련 이상

- 데이터 손실: 센서나 네트워크 장애로 아예 측정값이 누락된 경우
- 데이터 양 부족: 모델 · 분석이 필요로 하는 수량에 비해 충분히 데이터가 확보되지 않은 상태

이 두 현상은 그대로 완전성 저하로 이어질 수 있다. 즉, 필요 개수 대비 실제 수집 개수가 부족하면 완전성이 낮아진다.

3) 품질 측정

데이터 세트의 완전성 = A / B

A = 실제로 수집된 데이터 개수
B = 이론상 혹은 요구사항상 예상·필요 데이터 개수

3.4 일관성

1) 정의

일관성(consistency)은 센서 데이터가 단일 센서 내 또는 복수 센서 간에 사전에 정의된 규칙이나 논리에 어느 정도 부합하는지를 나타냄.

- 예: 센서 한 대가 시간에 따라 변하는 패턴이 규칙적으로 이어져야 하는데, 특정 시점에 전혀 맞지 않는 변동이 발생하거나, 복수 센서 간 합산 · 비교 규칙이 지켜지지 않는 경우 일관성이 깨진다.

2) 관련 이상

- 단일 센서 관점: 편이, 급락/급상승, 값 고정, 범위 진동 등은 시간 흐름 속에서 갑작스럽게 패턴을 깨뜨리거나 값 범위를 바꾸는 유형으로 일관성 저하 유발

- 복수 센서 관점: 이질성, 규칙 위반(예: A + B = C라는 공정 규칙 불일치)은 센서 간 상호 모순되는 값이 나타나 일관성을 저해

3) 품질 측정

데이터 세트의 일관성 = A / B

A = 주어진 규칙(혹은 패턴)에 부합하는 데이터 값의 개수
B = 전체 데이터 개수

센서 간 일관성을 평가할 때는 두 센서 혹은 여러 센서에서 동시에 측정된 값들이 사전에 정한 관계(예: 동등, 합, 차 등)에 맞는지 비율로 계산 가능.

3.5 정밀성[1]

정밀성(precision)은 표현 정밀성(representational precision)과 측정 정밀성(measurement precision)으로 구분하여 설명한다.

3.5.1 표현 정밀성

1) 정의

표현 정밀성(representational precision)은 센서가 데이터를 기록 및 표현할 때 사용되는 소수점 자릿수 혹은 해상도가 얼마나 충분히 유지되는가를 나타냄.

2) 관련 이상

해상도 불일치가 대표적.

- 예: 온도 센서가 99.999℃까지는 소수 셋째 자리까지 기록하는데, 100℃ 이상이면 소수 둘째 자리까지만 기록 → 해상도 불일치로 인해 데이터 정밀성이 떨어질 수 있음.

3) 품질 측정

데이터 세트의 표현 정밀성 = A / B

A = 요구된 소수점 자릿수를 만족하는 데이터값 개수
B = 전체 데이터값 개수

3.5.2 측정 정밀성

1) 정의

측정 정밀성(measurement precision)은 센서가 동일하거나 유사한 조건에서 반복 측정했을 때, 값들이 얼마나 좁은 분산 범위 안에 존재하는가를 의미함.

---

[1] ISO/IEC Guide 99와 ISO 5725-1은 표현 정밀성을 정밀성의 한 유형으로 인정하지 않는다. ISO/IEC Guide 99에서 정밀성은 동일하거나 유사한 대상을 특정 조건하에서 반복 측정하여 얻은 측정량 값들 간의 일치 정도(closeness of agreement)로 정의된다. 이러한 특성은 일반적으로 특정 측정 조건에서의 표준편차, 분산 또는 변동계수와 같은 불정밀성의 척도(measures of imprecision)에 의해 수치적으로 표현된다.

2) 관련 이상

주로 스파이크나 노이즈 등이 측정값의 분산을 크게 만들거나 이상을 생성하여 정밀성을 해칠 수 있다.

3) 품질 측정

데이터 세트의 측정 정밀성 = A / B

A = 사전에 설정된 분산(또는 표준편차, 혹은 특정 임계 구간) 이내에 속하는 데이터값 개수
B = 전체 데이터값 개수

3.6 적시성

1) 정의

적시성(timeliness)은 센서 데이터가 "요구되는 시간 내에 기록·전달되어, 시계열적 가치를 유지하는 정도"

- 예: 1초 단위로 제어해야 하는 로봇 시스템에서, 1초마다 센서 데이터가 도착하지 않고 지연이 발생하면 적시성이 크게 떨어짐.

2) 관련 이상

- 지연: 기록·전달이 지연되어 버퍼링이 누적되거나, 후속 처리(제어·분석 등)에 영향을 미치는 경우
- 주기 불일치: 측정 주기가 불규칙하게 바뀌어, 시간 축이 고르지 않은 간격으로 측정값을 생산
- 잘못된 시간정보: 타임스탬프가 잘못 기록됨으로써 실제 발생 시점과 데이터가 어긋남
- 시간정보 불일치(복수 센서): 센서 A와 B가 같은 물리적 이벤트를 서로 다른 시간에 발생한 것처럼 기록하는 현상

3) 품질 측정

데이터 세트의 적시성 = A / B

A = 타임스탬프가 요구 범위(지연 한계, 동기화 한계 등) 안에 들어오는 데이터값 개수
B = 전체 데이터값 개수

A/B가 높을수록 적시성이 우수함을 의미.

3.7 품질 특성과 ISO 8000-220, 230과의 연계

앞에서 정의한 다섯 가지 품질 특성 (정확성, 완전성, 일관성, 정밀성, 적시성)은 센서 데이터의 활용 가치를 결정짓는 핵심 지표다. 그리고 센서 데이터의 이상 유형은 센서 데이터 품질 특성을 저해하는 주요 요소이다.

ISO 8000-220에서는 위에서 정의한 품질 특성과 데이터 이상을 정량적으로 측정하기 위한 품질 측정 지표(quality measure)와 방법론(measurement function)을 제시한다. (부록 B 참조)

–예: 오프셋으로 인한 부정확성(inaccuracy due to offset) = (오차값이 허용범위 초과인 샘플 수) / (전체 샘플 수)

–스파이크 빈도, 노이즈 수준(분산·표준편차), 데이터 누락률 등

ISO 8000-230에서는 이러한 품질 특성과 측정 지표를 기반으로, 이상을 검출하고 데이터 정제 (오류값 보정, 제거, 대체 등)을 실행하여 센서 데이터 품질을 개선하는 지침을 제공한다.

예: 정확성 향상을 위해 오프셋이나 드리프트를 보정하고, 완전성 확보를 위해 데이터 손실에 대한 보간법 등이 수행될 수 있음.

구체적인 정제 프로세스는 4장, 이상 탐지(detection) 및 보정(repair) 방법은 부록 C, 데이터 정제 사례는 부록 D를 참조하기 바란다.

4 센서 데이터 정제를 위한 프로세스

ISO 8000-230은 센서 데이터에서 발생하는 다양한 이상을 탐지(detection)하고, 이를 삭제, 대체, 보간법 등의 방식으로 보정(repair)하여 전체 데이터 품질을 개선하는 절차인 센서 데이터 정제(cleansing) 프로세스를 다룬다.

본 장에서는 프로세스 체계 및 성격, 그리고 프로세스 활동 상세를 설명하며, ISO 8000-210(센서 데이터 품질 특성), ISO 8000-220(센서 데이터 품질 측정)을 바탕으로 실제 현장에서 센서 데이터를 정제하여 품질을 향상시키는 실무적 가이드를 제시한다.

4.1 프로세스 체계 및 성격

센서 데이터 정제 프로세스는 Plan-Do-Check-Act(이하 PDCA) 개념을 바탕으로 설계되었다. 즉, ISO 8000-61 [1]에서 데이터 품질 관리 프로세스를 정의하는 데 사용된 PDCA 개념이 데이터 정제 프로세스에 적용된다. 즉, 이 프로세스는 다음과 같은 활동으로 구성됨: 품질 측정 계획 수립(Plan), 데이터 품질 측정(Do 및 Check), 데이터 품질 개선(Act). 또한, 계획이 제공된 후에는 측정(Do와 Check) 및 개선(Act) 활동이 반복적으로 수행되어 센서 데이터가 품질 요구사항을 충족하는지 확인한다.

이 프로세스는 실시간 처리(또는 온라인 모드)가 아닌 사후처리(또는 오프라인 모드)를 위해 설계되었다. 센서 데이터는 실시간 스트림 형태로 수집되며 데이터 양이 매우 방대하기 때문에 이를 정제하는 데 시간이 소요된다. 따라서 신속한 의사결정이 필요한 환경에서는 실시간 데이터 정제가 현실적이지 않다. 실시간 데이터 정제는 데이터 이상이 이미 알려져 있으며 확인이나 검증할 필요가 없는 제한된 환경에서만 수행될 수 있다.

이 프로세스는 ISO/IEC/IEEE 31320-1 [2]에서 정의된 IDEF0(기능 모델링을 위한 통합 정의) 기능 모델로 표현된다. 이 모델은 프로세스를 계층적 활동으로 분해하여 어떤 활동이 수행되고 어떻게 수행되는지 보여준다. 각 활동의 입력, 출력, 제어, 및 수단을 명확히 보여줌으로써 프로세스의 분석 및 설계에 도움을 준다[2].

[2] 기능 모델은 모델 이름으로 식별되며, IDEF0 상자는 상자 이름으로, IDEF0 화살표 세그먼트는 화살표 라벨로 식별된다. 식별자는 제목 대소문자 표기법으로 작성되며, 즉 각 단어의 첫 글자가 대문자로 표기된다. 기능 모델의 표기법에 대한 자세한 내용은 ISO/IEC/IEEE 31320-1 [2]을 참조.

프로세스의 최하위 수준에 있는 각 활동(Activity)은 다음과 같다[3]:

- 제목(title): 활동에 대한 설명적인 표제
- 목적(purpose): 활동으로 달성하고자 하는 목표를 기술
- 작업(tasks): 활동의 목적 달성에 기여하기 위해 요구되거나 권장되거나 허용되는 행위들
- 입력(inputs): 활동에 필요한 입력 항목들
- 출력(outputs): 활동에 의해 산출되는 산물, 결과 또는 서비스
- 통제(controls): 활동이 올바른 출력을 생성하기 위해 필요한 조건 또는 제약사항
- 수단(mechanisms): 활동이 입력을 출력으로 변환하기 위해 사용하는 수단

4.2 프로세스 활동 상세

4.2.1 센서 데이터 정제 수행 (A0)

센서 데이터 정제(Perform Sensor Data Cleansing) 프로세스의 기능적 모델은 '센서 데이터 정제 수행'을 나타내는 A-0 컨텍스트 다이어그램으로 표현된다(그림 4-1참조).

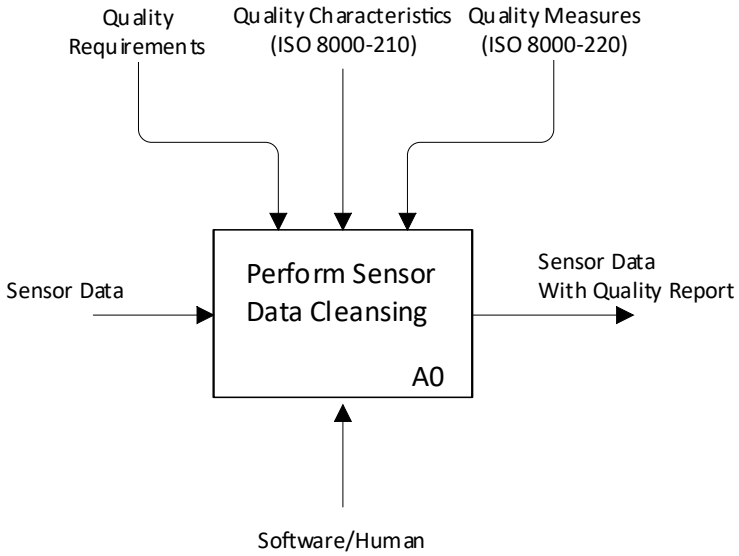


그림 4: 센서 데이터 정제 수행을 위한 A-0 컨텍스트 다어어그램 (모델 다이어그램 A0)

이 프로세스는 데이터 분석이나 활용 이전에 센서 데이터의 품질을 향상시키기 위해 데이터 정제를 수행하게 된다. 이 프로세스는 센서 데이터를 입력하고, 품질 요구사항, ISO 8000-210에서 정의한 품질 특성, ISO 8000-220에서 정의한 품질 측정을 제어 자료로 고려함으로써, 품질 보고서를 포함한 센서 데이터를 출력으로 제공한다. 이때 소프트웨어 및 인간이 메커니즘으로 프로세스를 수행하게 된다.

[3] 이러한 요소들은 ISO/IEC TR 24774:2010 [43], ISO/IEC/IEEE 24774:2021 [42]의 프로세스 설명 요소들과 ISO/IEC/IEEE 31320-1 [2]의 기능 모델 요소를 활동 정의에 맞게 수정.

이 프로세스는 그림 4-2와 같이 세 가지 활동으로 구성된다[4]:

- 측정 계획 수립 (Prepare Measurement Plan, A1)
- 데이터 품질 측정 (Measure Data Quality, A2)
- 데이터 품질 개선 (Improve Data Quality, A3)

4.2.2 측정 계획 수립 (A1)

측정계획수립 활동은 품질 요구사항, 품질 특성, 품질 측정지표 및 센서 데이터를 기반으로 센서 데이터 품질을 측정하기 위한 계획을 준비하는 것을 목적으로 한다.

그림 4-2과 같이, 이 활동은 세 가지 하위 활동으로 구성된다:

- 데이터 품질 목표 수립(Establish Data Quality Goal, A11)
- 데이터 프로파일링 수행(Perform Data Profiling, A12)
- 측정 계획 개발(Develop Measurement Plan, A13).

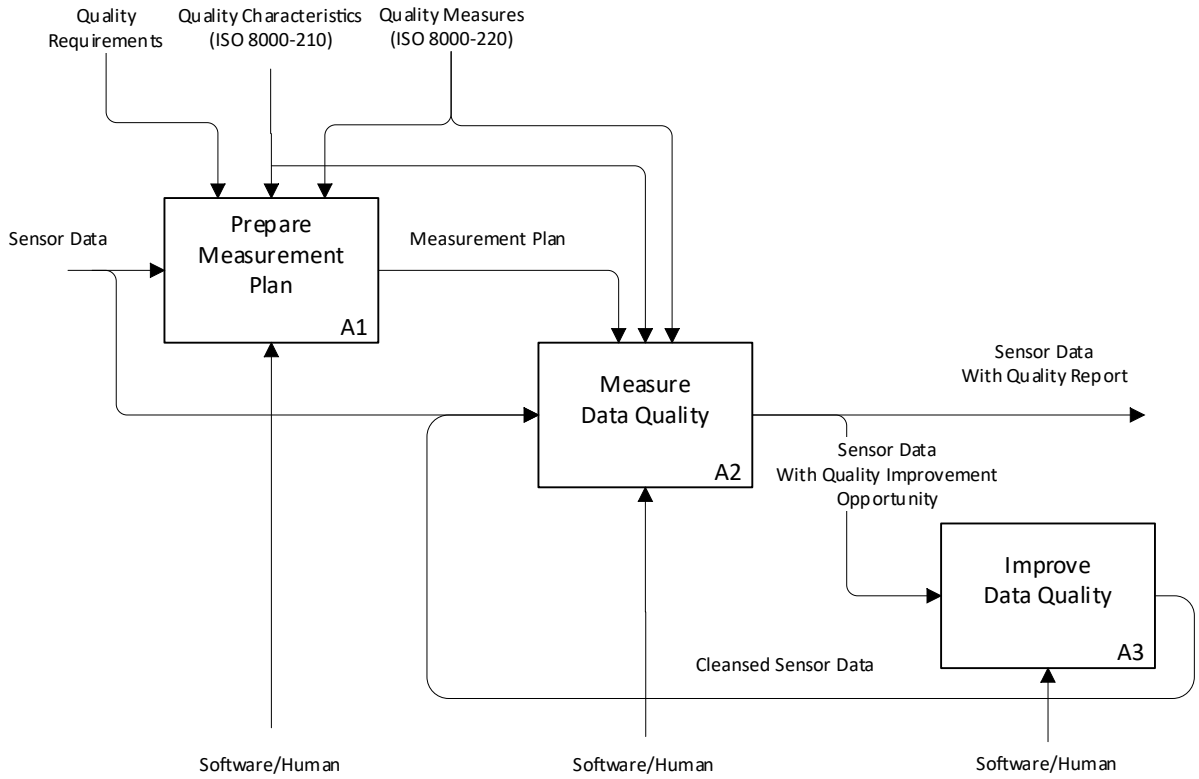


그림 5: 센서 데이터 정제 수행 (모델 다이어그램 A0)

4.2.2.1 데이터 품질 목표 수립 (A11)

1) 목적

- 센서 데이터의 품질 요구사항을 반영하는 데이터 품질 관련 목표를 결정

[4] 그림 4-1의 하위 다이어그램

2) 작업:

- 이해관계자로부터 데이터 품질 요구사항을 수집
- 데이터 품질 요구사항을 기반으로 달성해야 할 목표를 결정

3) 입력

- 센서 노드에서 수집된 센서 데이터

4) 출력

- 관심 있는 품질 특성의 품질 측정 수준 등 데이터 품질 요구사항으로 표현된 데이터 품질 목표

5) 통제

- ISO 8000-210에서 정의된 품질 요구사항, 품질 특성 및 해당 데이터 이상
- ISO 8000-220에서 정의된 품질 측정지표

6) 수단

- 소프트웨어/인간

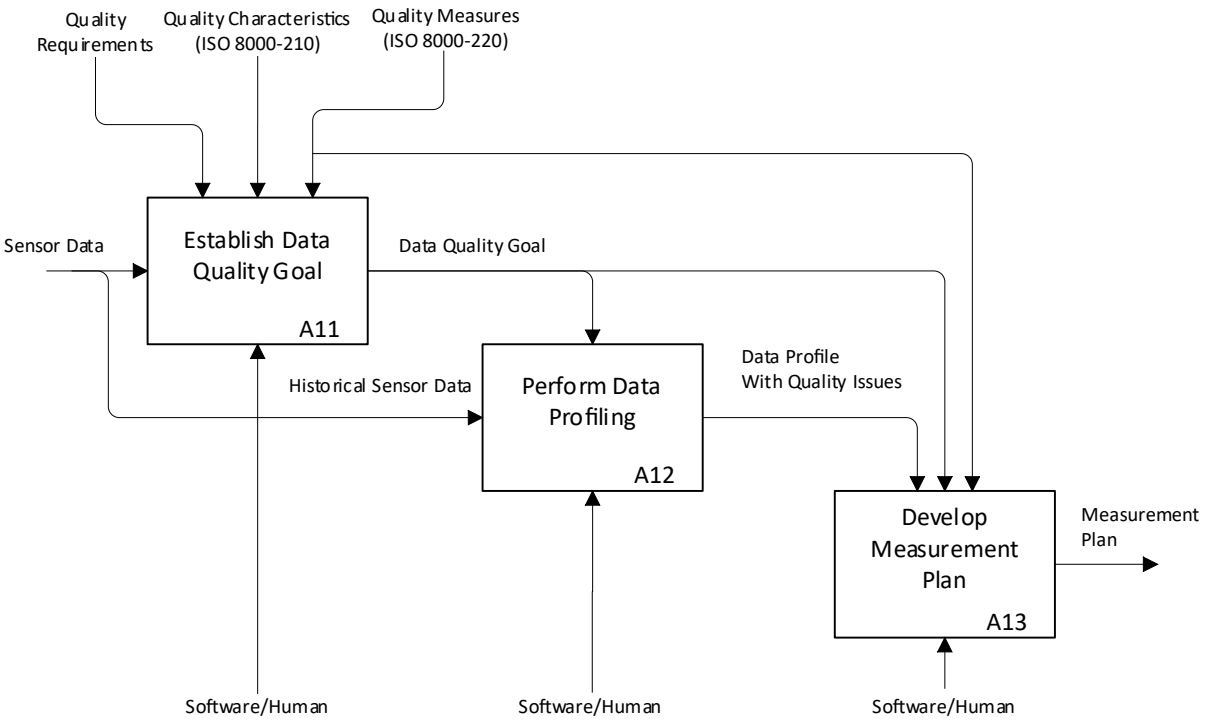


그림 6: 측정 계획 수립(모델 다이어그램 A1)

4.2.2.2 데이터 프로파일링 수행 (A12)

1) 목적

- 과거의 센서 데이터를 획득하고 이에 대한 데이터 프로파일링을 수행
- 관련 센서 데이터의 과거 발생 집합으로부터 센서 데이터의 프로파일과 데이터 품질 문제를 추출

2) 작업
   - 과거 센서 데이터를 수집
   - 센서 데이터에 대한 데이터 프로파일링[5]을 수행
3) 입력
   - 과거 센서 데이터
4) 출력
   - 품질 문제가 포함된 데이터 프로파일
5) 통제
   - 데이터 품질 목표
6) 수단
   - 통계적, 수학적, 데이터 학습 기법을 제공하는 소프트웨어, 또는 정보를 대화형/수동으로 입력하는 인간

4.2.2.3 측정 계획 개발 (A13)

1) 목적
   - 참조 데이터 패턴에 따라 센서 데이터의 품질을 측정하는 데 사용될 방법, 절차, 기준 및 근거를 포함하는 측정 계획을 수립
2) 작업
   - 데이터 품질을 측정하기 위한 방법과 절차를 정의
   - 데이터 품질을 평가하는 데 필요한 기준과 정보를 결정
3) 입력
   - 없음
4) 출력
   - 측정 계획
5) 통제
   - 데이터 품질 목표, 품질 문제가 포함된 데이터 프로파일
   - ISO 8000-220에서 정의된 품질 측정지표
6) 수단
   - 소프트웨어/인간

[5] ISO/TS 8000-81 [3]을 참조

4.2.3 데이터 품질 측정 (A2)

데이터 품질 측정(Measure Data Quality) 활동은 수립된 측정 계획에 따라 센서 데이터의 이상 탐지 모델과 품질 측정값을 도출하고, 품질 개선의 기회를 식별하는 것을 목적으로 한다.

그림 4-4와 같이, 이 활동은 세 가지 하위 활동으로 구성된다:

- 이상 탐지 모델 도출 (Derive Anomaly Detection Model, A21)
- 품질 개선 기회 발견 (Find Quality Improvement Opportunity, A22)
- 품질 결과 보고 (Report Quality Result, A23)

4.2.3.1 이상 탐지 모델 도출 (A21)

1) 목적
   - 센서 데이터의 데이터 패턴을 분석하고, 데이터 이상을 탐지할 수 있는 모델을 결정
2) 작업
   - 데이터 패턴 분석
   - 이상 탐지 모델[6] 결정
3) 입력
   - 센서 데이터
4) 출력
   - 이상 탐지 모델
5) 통제
   - 측정 계획
6) 수단
   - 소프트웨어/인간

[6] 이상 탐지 모델에 대해서는 B.1절을 참조. 이 모델은 이상 유형을 식별하거나 센서 데이터에 포함된 이상 값을 탐지하는 기능을 가진다.

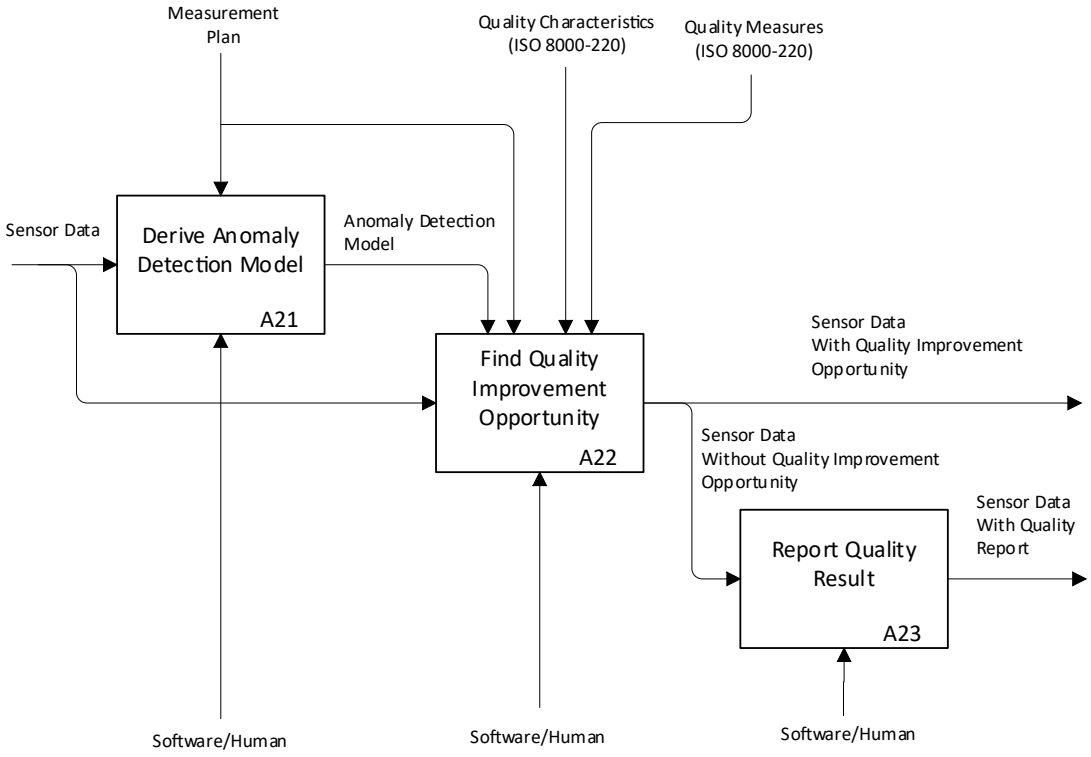


그림 7: 데이터 품질 측정 (모델 다이어그램 A2)

4.2.3.2 품질 개선 기회 발견 (A22)

1) 목적

- 이상 탐지 모델에 기반하여 품질 측정값을 평가하고, 데이터 이상을 수정함으로써 센서 데이터 품질을 개선 기회의 발견

2) 작업

- 품질 특성별 품질 측정값 평가
  - i) ISO 8000-220에 정의된 품질 특성별 품질 측정값을 측정
  - ii) 만약 품질 요구사항을 충족한다면, 센서 데이터는 품질 개선이 필요하지 않으므로 작업을 종료
  - iii) 그렇지 않은 경우, 데이터 품질에 영향을 주는 데이터 이상에 대해 다음 작업을 추가로 수행
- 이상별 품질 측정값 평가
  - i) 센서 데이터에 포함된 데이터 이상을 탐지하고, ISO 8000-220에 정의된 이상별 품질 측정값을 측정
  - ii) 만약 품질 특성별 품질 측정값을 개선하거나 이상별 품질 측정값을 감소시킬 수 있도록 수정 가능한 데이터 이상이 존재한다면, 해당 센서 데이터는 품질 개선 기회가 있는 것으로 간주
  - iii) 그렇지 않으면, 품질 개선 기회가 없는 센서 데이터로 간주

3) 입력

- 센서 데이터

4) 출력
   - 품질 개선 기회가 있는 센서 데이터: 센서 데이터 품질을 개선하기 위해 수정 가능한 데이터 이상이 식별된 경우
   - 품질 개선 기회가 없는 센서 데이터: 품질 요구사항을 충족하여 개선이 필요하지 않거나, 품질 개선에 기여할 데이터 이상이 식별되지 않은 경우
5) 통제
   - 이상 탐지 모델, 측정 계획, ISO 8000-220에 정의된 품질 특성과 품질 측정값
6) 수단
   - 소프트웨어/인간

4.2.3.3 품질 결과 보고 (A23)

1) 목적
   - 센서 데이터의 품질 결과를 보고
2) 작업
   - 품질 요구사항, 문제점, 개선 사항 등을 포함한 품질 정보를 수집
   - 품질 개선 노력을 반영한 보고서 작성
3) 입력
   - 품질 개선 기회가 없는 센서 데이터(품질 요구사항을 충족하거나, 개선 가능한 데이터 이상이 식별되지 않은 경우)
4) 출력
   - 품질 정보(품질 요구사항, 정제를 통한 개선, 문제점 등)를 포함하는 품질 보고서가 있는 센서 데이터. 이 데이터는 다음 두 가지 범주 중 하나에 속한다:
     i) 품질 요구사항을 충족하여 데이터 분석 또는 활용에 사용할 수 있는 센서 데이터
     ii) 품질 요구사항을 충족하지 못하며, 더 이상 품질 개선이 불가능하여 데이터 분석 또는 활용에 사용할 수 없는 센서 데이터 → 이 경우, 센서 데이터는 폐기되거나 데이터 품질 저하의 원인을 심층 분석하는 대상으로 전환
5) 통제
   - 없음
6) 수단
   - 소프트웨어/인간

4.2.4 데이터 품질 개선 (A3)

데이터 품질 개선 (Improve Data Quality) 활동은 확인된 데이터 품질 개선 기회를 바탕으로 센서 데이터를 정제하여 데이터 품질을 개선하고, 정제된 센서 데이터를 제공하는 것을 목적으로 한다.

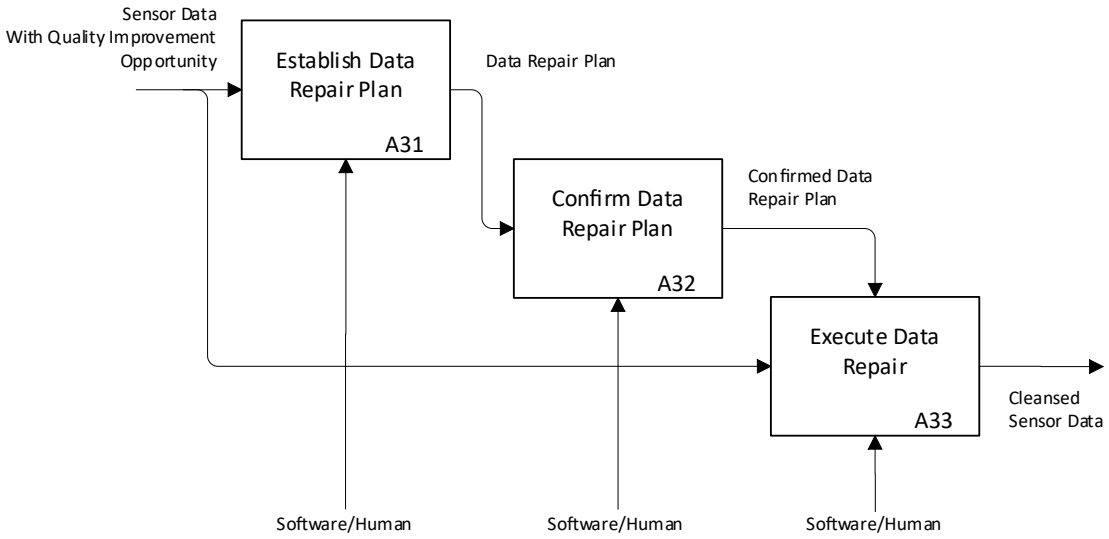


그림 8: 데이터 품질 개선(모델 다이어그램 A3)

그림 4-5와 같이, 이 활동은 세 가지 하위 활동으로 구성된다:

- 데이터 보정 계획 수립 (Establish Data Repair Plan, A31)
- 데이터 보정 계획 확인 (Confirm Data Repair Plan, A32)
- 데이터 보정 실행 (Execute Data Repair, A33)

4.2.4.1 데이터 보정 계획 수립 (A31)

1) 목적
   - 확인된 품질 개선 기회를 바탕으로 센서 데이터를 정제하기 위한 구체적인 실행 계획을 수립
2) 작업
   - 데이터 이상값을 보정할 수 있는 대체 방법을 나열
   - 데이터 보정 계획을 결정
3) 입력
   - 품질 개선 기회가 있는 센서 데이터
4) 출력
   - 데이터 보정 계획
5) 통제
   - 없음
6) 수단
   - 소프트웨어/인적 자원

4.2.4.2 데이터 보정 계획 확인 (A32)

1) 목적
   - 수립된 데이터 보정 계획에 대해 다양한 이해관계자의 확인을 받고 이를 확정

- 데이터 보정으로 인해 발생할 수 있는 위험과 문제에 대해 이해관계자들이 충분히 인지하고 동의하도록 보장

2) 작업:
   - 데이터 보정 계획에 대해 이해관계자의 의견을 수집
   - 데이터 보정 우선순위를 포함하여 실행 가능한 데이터 보정 계획을 확정
3) 입력
   - 데이터 보정 계획
4) 출력
   - 확정된 데이터 보정 계획
5) 통제
   - 없음
6) 수단
   - 소프트웨어/인적 자원

4.2.4.3 데이터 보정 실행 (A33)

1) 목적
   - 수립된 데이터 보정 계획을 실제로 실행하여 정제된 센서 데이터를 제공
2) 작업
   - 데이터 보정[7] 실행 계획을 구체화한다.
   - 데이터 보정을 실행하고 결과를 확인한다.
3) 입력
   - 품질 개선 기회가 있는 센서 데이터
4) 출력
   - 정제되었음을 명시한 정제된 센서 데이터
5) 통제
   - 확정된 데이터 보정 계획
6) 수단
   - 데이터 이상 보정 방법을 포함한 소프트웨어/인적 자원

[7] 데이터 보정 실행 방법에 대해서는 부록 C.2를 참조

5 사용자 가이드라인의 효과적 활용을 위한 요구사항

1) 센서 데이터를 정제하기 위해서는 센서 데이터는 다음 요구사항을 충족해야 한다:
   - 식별 가능(identifiable).
   - 사전에 정의된 데이터 형식(data formats)을 준수.
   - 쉽게 접근가능(accessible)하고, 이해가능(understandable)해야함.
2) 센서 데이터를 잘 정제하기 위해서는 다음 표준을 상호 연계하여 정제 프로세스를 진행할 것을 권고한다:
   - ISO 8000-210에서 규정한 데이터 품질 특성 및 이상
   - ISO 8000-220에서 규정한 데이터 품질 측정지표
   - ISO 8000-230에서 규정한 데이터 정제 가이드라인
3) 조직 내부 절차, 이해관계자 승인
   - 데이터 정제는 '원본 데이터 수정'이 수반될 수 있으므로, 내부 규정 및 이해관계자 동의가 필요하다.
   - 특히, 보안, 개인정보 보호, 규제 준수 등의 추가 요구 사항이 존재할 수 있다.
4) 실시간(Online) 또는 오프라인(Offline) 처리:
   - 본 문서는 오프라인 데이터 정제 프로세스를 제시하고 있다.
   - 실시간 모니터링 및 제어가 필요한 영역에서는 이 문서에서 제시한 정제 프로세스를 변형하여 사용할 것을 권고한다.

6 참고 문헌

[1] ISO, "ISO 8000-61, Data quality — Part 61: Data quality management: Process reference model".

[2] ISO, "ISO/IEC/IEEE 31320-1, Information technology — Modeling Languages — Part 1: Syntax and Semantics for IDEF0".

[3] ISO, "ISO/TS 8000-81, Data quality — Part 81: Data quality assessment: Profiling," ISO.

[4] ISO, "ISO 18113-1, In vitro diagnostic medical devices — Information supplied by the manufacturer (labelling) — Part 1: Terms, definitions, and general requirements," ISO, 2022.

[5] ISO, "ISO/IEC 29182-2, Information technology — Sensor networks: Sensor Network Reference Architecture (SNRA) — Part 2: Vocabulary and terminology," ISO/IEC, 2013.

[6] ISO, "ISO 8000-8, Data quality — Part 8: Information and data quality: Concepts and measuring," ISO.

[7] ISO, "ISO 9000, Quality management systems — Fundamentals and vocabulary," ISO, 2015.

[8] ISO, "ISO 8000-2, Data quality — Part 2: Vocabulary," ISO.

[9] ISO, "ISO 13008, Information and documentation — Digital records conversion and migration process," 2022.

[10] ISO, "ISO/IEC 25012, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model," ISO/IEC, 2008.

[11] X. C. ZHAI, "Outlier Detection in Wireless Sensor Networks Using Dynamic Threshold," %1 *Electronic technology, 2015, 28(2), 18-21.*.

[12] G. KERSCHEN, P. D. BOE, J. GOLINVAL 그리고 K. Worden, "Sensor validation using principal component analysis," %1 *Smart Materials and Structures*, 2005.

[13] H. P. VINUTHA, B. POORNIMA 그리고 B. M. SAGAR, "Detection of outliers using interquartile range technique from intrusion dataset," %1 *Information and decision sciences: Proceedings of the 6th international conference on FIC TA*, 2018.

[14] Z. CHENG, C. ZOU 그리고 J. DONG, "Outlier detection using isolation forest and local outlier factor," %1 *Proceedings of the conference on research in adaptive and convergent systems*, 2019.

[15] E. J. JAMSHIDI, Y. YUSUP, J. S. KAYODE 그리고 M. A. KAMARUDDIN, "Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature," *Ecological Informatics,* 제 69, 번호: 101672, 2022.

[16] J. R. JIANG, J. B. KAO 그리고 Y. L. LI, "Semi-supervised time series anomaly detection based on statistics and deep learning," *Applied Sciences,* 제 11(15), 번호: 6698, 2021.

[17] O. O. ATIENZA, L. C. TANG 그리고 B. W. ANG, "A SPC Procedure for Detecting Level Shifts of Autocorrelated Processes," *Journal of Quality Technology,* 제 30(4), pp. 340-351, 1998.

[18] H. JIN, G. YIN, B. YUAN 그리고 F. JIANG, "Bayesian hierarchical model for change point detection in multivariate sequences," *Technometrics,* 제 64(2), pp. 177-186, 2021.

[19] P. CLOSAS 그리고 C. FERNANDEZ-PRADES, "Particle filtering with adaptive number of particles," %1 *Aerospace Conference*, 2011.

[20] H. Q. ZHANG 그리고 Y. YAN, "A wavelet-based approach to abrupt fault detection and diagnosis of sensors," *IEEE Transactions on Instrumentation and Measurement,* 제 50(5), pp. 1389-1396, 2001.

[21] Y. M. WANG, R. G. LI 그리고 H. CHAI, "Detection of Advanced Pulse Compression Noise Based on FRFT," %1 *IEEE 5th International Conference on Electronic Information and Communication Technology (ICEICT)*, 2022.

[22] C. ROMESIS 그리고 K. MATHIOUDAKIS, "Setting Up of a Probabilistic Neural Network for Sensor Fault Detection Including Operation with Component Faults," *Gas Turbines Power,* 제 125(3), p. 634–641, 2003.

[23] V. KOZITSIN, I. KASTER 그리고 D. LAKONTSEV, "Online forecasting and anomaly detection based on the ARIMA model," *Applied Sciences,* 제 11(7), 번호: 3194, 2021.

[24] D. XU, Y. WANG, Y. MENG 그리고 Z. ZHANG, "An improved data anomaly detection method based on isolation forest," %1 *2017 10th international symposium on computational intelligence and design*, 2017.

[25] M. AMER, M. GOLDSTEIN and S. ABDENNADHER, M. GOLDSTEIN 그리고 S. ABDENNADHER, "Enhancing one-class support vector machines for unsupervised anomaly detection," %1 *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, 2013.

[26] F. ESMAEILI, E. CASSIE, H. P. T. NGUYEN 그리고 N. O. PLA, "Anomaly detection for sensor signals utilizing deep learning autoencoder-based neural networks," *Bioengineering,* 제 10(4), 번호: 405, 2023.

[27] Y. WANG, M. PERRY, D. WHITLOCK 그리고 J. W. SUTHERLAN, "Detecting anomalies in time series data from a manufacturing system using recurrent neural networks," *Journal of Manufacturing Systems,* 제 62, pp. 823-834, 2022.

[28] I. PRATAMA, A. E. PERMANASARI, I. ARDIYANTO 그리고 R. INDRAYANI, "A Review of Missing Values Handling Methods on Time-Series Data," %1 *International Conference on Information Technology Systems and Innovation (ICITSI)*, 2016.

[29] F. ZHANG, X. TANG, A. TONG, B. WANG 그리고 J. WANG, "An Automatic Baseline Correction Method Based on the Penalized Least Squares Method," *Sensors,* 제 20(7), 번호: 2015, 2020.

[30] F. YE, J. CHEN 그리고 Y. B. LI, "Improvement of DS Evidence Theory for Multi-Sensor Conflicting Information," *Symmetry,* 제 69, 2017.

[31] D. WATZENIG 그리고 G. STEINER, "Offset compensation for capacitive angular position sensors by evaluating the Kalman filter innovation sequence," %1 *IEEE International Conference on Industrial Technology*, 2004.

[32] M. MATERASSI, L. ALFONSI, G. DE FRANCESCHI, V. ROMANO, C. MITCHELL 그리고 P. SPALLA, "Detrend effect on the scalograms of GPS power scintillation," *Advances in Space Research,* 제 43(11), pp. 1740-1748, 2009.

[33] R. YE, H. HE, P. ZHENG, M. XU 그리고 L. A. WANG, "Spike Removal Algorithm Based on Median Filter and Statistic for Raman Spectra," *SPECTROSCOPY AND SPECTRAL ANALYSIS,* 제 42(10), pp. 3174-3179, 2022.

[34] J. LIN, X. SUN, J. WU, S. C. CHAN 그리고 W. XU, "Removal of power line interference in EEG signals with spike noise based on robust adaptive filter," %1 *IEEE Region 10 Conference (TENCON)*, 2016.

[35] G. VENERI, p. FEDERIGHI, f. ROSINI, A. FEDERICO 그리고 A. RUFA, "Spike removal through multiscale wavelet and entropy analysis of ocular motor noise: A case study in patients with cerebellar disease," *Journal of neuroscience methods,* 제 196(2), pp. 318-326, 2011.

[36] S. GANDHAM 그리고 B. ANURADHA, "An iterative method of Ensemble Empirical Mode Decomposition for enhanced ECG signal denoising," %1 *International Conference on Wireless Communications, Signal Processing and Networking*, 2016.

[37] X. P. ZHANG, N. Q. HU, Z. CHENG 그리고 H. ZHONG, "Vibration data recovery based on compressed sensing," *Acta Physica Sinica,* 제 63(20), 2014.

[38] G. N. KAMM, "Computer Fourier-transform techniques for precise spectrum measurements of oscillatory data with application to the de Haas-van Alphen effect," *Journal of Applied Physics,* 제 49(12), 1978.

[39] D. FOLGADO, M. BARANDAS, R. MATIAS, R. MARTINS, M. CARVALHO 그리고 H. GAMBOA, "Time alignment measurement for time series," *Pattern Recognition,* 제 81, 2018.

[40] C. G. FANG 그리고 C. WANG, "Time Series Data Imputation: A Survey on Deep Learning Approaches," *Preprint submitted to Elsevier,* 2020.

[41] I. GOODFELLOW, A. J. POUGET, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE 그리고 Y. BENGIO, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems,* 제 27, 2014.

[42] ISO, "ISO/IEC/IEEE 24774, Systems and software engineering — Life cycle management — Specification for process description," ISO/IEC/IEEE, 2021.

[43] ISO, "ISO/IEC TR 24774, Systems and software engineering — Life cycle management — Guidelines for process description," ISO, 2010.

[44] Y. KATSUMOTO 그리고 Y. OZAKI, "Practical algorithm for reducing convex spike noises on a spectrum," *Applied spectroscopy,* 제 57(3), pp. 317-322, 2003.

부록

A. 용어 정의 (ISO 8000-210,220,230)

A.1 센서 데이터 관련 용어

1) 신호
   - 원문: signal
   - 정의: 측정대상(measurand)을 나타내며, 그와 기능적으로 연관된 양(量)
   - 출처: ISO 18113-1:2022 [4], 3.2.35, 수정 — 정의 measurement signal에서 비고 1과 비고 2는 삭제됨.

2) 센서
   - 원문: sensor
   - 정의: 자연 현상, 시스템 또는 인공적 활동의 속성을 관찰하고, 측정하고, 해당 측정값을 신호로 변환하는 장치
   - 참고 1: 센서는 단일 물리적 형태로 존재할 뿐만 아니라 가상 센서와 같이 변형된 형태로도 존재할 수 있다.
   - 출처: ISO/IEC 29182-2:2013 [5], 2.1.5, 수정 - "시스템"이 정의에 추가되고, "사람들"이 "인간"으로 변경되었으며, "물리적"이 정의에서 삭제되었다. 참고 1이 변경되었다.

3) 센서 네트워크
   - 원문: sensor network
   - 정의: 공간적으로 분산된 센서 노드로 구성된 시스템으로, 서로 상호 작용하고, 응용 프로그램에 따라서는 다른 인프라와도 상호 작용하여 주변 환경에서 추출한 정보를 획득, 처리, 전송 및 제공하는 주요 기능인 정보 수집 및 가능한 제어 기능을 수행한다.
   - 참고: 센서 네트워크의 구별되는 특징에는 광역 적용 범위, 무선 네트워크 사용, 목적의 유연성, 자체 구성, 개방성 및 여러 응용 프로그램에 대한 데이터 제공이 포함될 수 있다.

4) 센서 노드
   - 원문: sensor node
   - 정의: 최소한 하나의 센서와 선택적으로 통신 기능 및 데이터 처리 기능이 있는 구동기(액추에이터)를 포함하는 센서 네트워크 요소
   - 참고
     i) 여기에는 추가 응용 프로그램 기능이 포함될 수 있다.
     ii) 여러 센서로 구성된 하이브리드 센서는 여러 센서를 포함하는 센서 노드(3.1.3)로 간주된다.
   - 출처: ISO/IEC 29182-2:2013 [5], 2.1.8, 수정됨 - 정의에 항목에 대한 참고 사항 2가 추가되었다.

5) 센서 데이터
   - 원문: sensor data
   - 정의: 센서 노드에서 생성된 데이터(3.1.3)

- 참고: 센서 데이터(3.1.4)는 센서 신호로부터 변환된 디지털 값의 스트림과 각 센서의 식별 정보, 센서 노드(3.1.3)에서 수집한 데이터의 타임스탬프와 같은 정보로 구성된다.

6) 사물 인터넷
- 원문: internet of things
- 약어: IoT
- 정의: 센서, 소프트웨어 및 기타 기술이 내장되어 있어 데이터를 수집하고 교환할 수 있는 다양한 형태의 물리적 장치가 네트워크 형태로 연결된 인프라

A.2 데이터 품질 관련 용어

1) 데이터 이상
- 원문: data anomaly
- 정의: 데이터 집합의 데이터 항목이 데이터 집합의 예상 패턴과 다른 경우

2) 품질 특성
- 원문: quality characteristic
- 정의: 요구 사항과 관련된 객체의 고유한 특성
- 참고: ISO 8000-8 [6]은 데이터의 실용적 품질을 결정하는 품질 특성의 동의어로 품질 차원이라는 용어를 사용한다.
- 출처: ISO 9000:2015 [7], 3.10.2, 수정됨 - 참고가 추가되었다.

3) 데이터 정제
- 원문: data cleansing
- 정의: 데이터의 결함 및 오류를 탐지하고 수정하여 데이터 품질을 향상시키는 과정
- 참고
  i) ISO 8000-2 [8]에서 데이터 오류는 데이터 요구사항의 미준수로 정의되며, 데이터 부적합성과 동일시된다.
  ii) ISO 9000 [7]에서 결함은 의도된 또는 지정된 용도와 관련된 요구사항을 충족하지 못하는 것으로 정의된다.
  iii) ISO 8000-61 [1]에서 데이터 정제는 데이터 품질 개선의 하위 프로세스로 규정되어 있다.
- 출처: ISO 13008:2022 [9], 3.4, 수정 — "수정(또는 제거)"이 "수리"로 변경되었으며, 참고 1, 2 및 3이 추가되었다.

4) 데이터 프로파일링
- 원문: data profiling
- 정의: 감사 데이터 추출에 영향을 미치는 데이터 구조 및 시스템 규칙을 이해하기 위해 수행되는 활동

A.3 측정 관련 용어

1) 데이터 품질 측정지표 (품질 측정지표)

- 원문: data quality measure (quality measure)
- 정의: 데이터 품질 특성을 측정하는 결과로 값이 할당되는 변수
- 참고: ISO/IEC 25012:2008 [10]의 4.5절을 바탕으로 작성되었다.

2) 품질 측정 요소

- 원문: quality measure element
- 정의: 데이터 품질 측정을 구성하기 위해 사용되는 측정으로, 기본 측정값(base measurement) 또는 파생 측정값(derived measurement)일 수 있다.
- 참고: ISO/IEC 25012:2008 [10], 4.14에서 인용·수정하였다.

B. 데이터의 품질 측정지표

B.1 기본 원칙 및 가정

데이터 품질 측정지표 (data quality measure)는 데이터 품질을 측정하고 개선하기 위한 토대이다. 이러한 측정지표는 센서 네트워크와 사물인터넷(IoT)에 연결된 센싱 장치가 단일 이산 디지털 값의 스트림으로 기록한 센서 데이터를 포함하여 모든 종류의 데이터에 유용하다. 이러한 측정지표는 데이터 품질 특성 또는 데이터에 적용되는 데이터 이상을 정량화한다. (데이터 품질 특성인 정확성에 대한 데이터 품질 측정지표의 예는 표 B-1 참조).

각 측정지표는 적절한 품질 측정 요소를 조합하여 결과를 생성하는 측정 함수의 결과이다. 각 데이터 품질 특성이나 데이터 이상은 하나 이상의 데이터 품질 측정지표로 정량화될 수 있다.

표 B-1: 데이터 품질 특성(정확성)에 대한 데이터 품질 측정지표 예시

데이터 품질 측정지표	측정 함수	함수 내 품질 측정 요소
데이터 세트의 정확성	A / B	A = 정확한 것으로 평가된 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수

B.2 센서 데이터의 품질 측정지표

B.2.1 일반

센서 데이터가 충분한 품질을 갖추지 못하면, 데이터에는 일반적으로 하나 이상의 데이터 이상이 포함된다.

- 예시 1: 센서 데이터 집합의 정확성가 낮으면, 해당 집합에는 오프셋, 드리프트, 트림, 스파이크, 노이즈 등의 이상이 포함되는 경우가 많다.

따라서 데이터에서 이상을 제거하거나 줄이면 데이터 품질을 향상시킬 수 있다. 이를 지원하기 위해, 본 절은 ISO 8000-210에서 규정한 각 데이터 품질 특성과 해당 데이터 이상에 대한 품질 측정지표를 규정한다.

품질 측정지표는 다음 두 가지 유형 중 하나이다.

– 특정 품질 특성에 대한 것으로, 일정 기간 동안 수집된 데이터 값이 해당 품질 특성의 기준이나 요구사항을 얼마나 충족하는지를 나타내는 정도 또는 값

– 특정 데이터 이상에 대한 것으로, 일정 기간 동안 발견된 데이터 이상이 센서가 관측하는 속성의 참조(기대) 데이터 패턴에서 얼마나 벗어나는지를 나타내는 정도 또는 값

특정 품질 특성의 품질 측정지표가 낮으면, 그에 대응하는 데이터 이상들의 품질 측정지표는 일반적으로 상대적으로 높게 나타난다. 이러한 이상들은 해당 품질 특성에 영향을 미치는 요소이다. 따라서 데이터 제거, 보정(compensation) 등의 데이터 처리(data handling)를 통해 데이터 이상을 식별하고 그 측정지표를 낮춤으로써, 품질 특성의 측정지표를 향상시킬 수 있다.

1) 품질 측정지표는 [0,1] 범위의 이진 값, 백분율, 점수, 경계(bound), 빈도, 확률 등 다양한 유형의 값을 나타낼 수 있다.

2) 품질 측정지표는 의도된 사용 목적이나 사용자의 편의를 위해 수정될 수 있다.
   - 예시 2: 백분율 값을 요구사항을 충족하는 데이터 값의 개수로 변환한다.

- 예시 3: 품질 측정지표를 [0,1] 범위의 정규화(normalized) 값으로 변환한다.

3) 일부 데이터 이상은 둘 이상의 데이터 품질 특성에 영향을 미칠 수 있다. 각 이상과 품질 특성의 조합별로 해당 품질 측정지표는 다른 조합에서의 측정지표와 다를 수 있다.

본 문서는 다음 데이터 품질 특성에 대한 품질 측정지표를 규정한다.

- 정확성
- 완전성
- 일관성
- 정밀성
- 적시성

B.2.2 정확성에 대한 품질 측정지표

정확성 품질 측정지표는 특정 사용 맥락에서 센서 데이터가 의도된 개념 또는 사건의 속성의 참값을 얼마나 정확히 표현하는지를 나타낸다.

이러한 측정지표에는 오프셋, 드리프트, 트림, 스파이크, 노이즈 등 정확성에 영향을 미칠 수 있는 데이터 이상에 대한 측정지표도 포함된다.

품질 특성 및 데이터 이상에 대한 측정지표는 표 B-2 에 요약되어 있다.

표 B-2: 정확성 및 관련 데이터 이상에 대한 품질 측정지표

구분	품질 측정지표	설명	측정 함수
품질 특성 기준	데이터 세트의 정확성	특정 사용 맥락에서 센서 데이터 집합이 의도된 속성의 진정한 값을 올바르게 표현하는 값으로 구성된 정도	X = A / B A = 정확한 것으로 평가된 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
이상 기준	오프셋으로 인한 부정확성	센서 데이터의 오프셋으로 인해 값이 의도된 속성을 올바르게 표현하지 못하는 정도	X = A / B A = 부정확하다고 평가된 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	드리프트로 인한 부정확성	드리프트가 값의 참 표현을 왜곡하는 정도	X = A / B A = 부정확하다고 평가된 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	트림으로 인한 부정확성	센서 측정 능력의 한계로 값이 절단 (truncation)되어 진정한 값을 표현하지 못하는 정도	X = A / B A = 적용 가능한 정상 또는 사전 정의된 패턴 범위를 벗어난 것으로 평가된 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	스파이크로 인한 부정확성[8]	센서 데이터가 정상 또는 사전 정의된 패턴과 일치하지 않는 정도	X = A / B A = 정상 또는 사전 정의된 패턴 범위를 벗어난 데이터 값 개수 B = 평가된 전체 데이터 값의 개수

[8] "스파이크로 인한 부정확성"은 데이터가 정상 또는 사전 정의된 데이터 패턴에서 벗어나는 경우를 나타낸다. 이러한 이상은 열악한 환경, 센서 결함, 시스템 오류 등 다양한 원인으로 발생할 수 있으며, 이러한 원인은 데이터 품질 개선 조치로 해결할 수 있다.

구분	품질 측정지표	설명	측정 함수
	노이즈로 인한 부정확성	센서 데이터 값이 참 표현에서 무작위로 변동하는 정도	X = A / B A = 값과 의도된 속성의 참 표현 간의 적용 가능한 정상 또는 사전 정의된 변동 범위를 벗어난 것으로 평가된 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수

B.2.3 완전성을 위한 품질 측정지표

완전성 품질 측정지표는 특정 사용 맥락에서 센서 데이터 집합이 기대되는 모든 값을 포착하고 있는 정도를 나타낸다. 이 품질 측정지표는 또한 데이터 손실이나 데이터 양 부족과 같은 데이터 이상을 포함하며, 이러한 요소들은 완전성에 영향을 줄 수 있다. 품질 특성 및 데이터 이상에 특화된 품질 측정지표는 표 B-3 에 요약되어 있다.

표 B-3: 완전성 및 관련 데이터 이상에 대한 품질 측정지표

구분	품질 측정지표	설명	계산식
품질 특성 기준	데이터 세트의 완전성	특정 사용 맥락에서 센서 데이터 집합이 기대되는 모든 값을 포착하고 있는 정도	X = A / B A = 실제로 포착된 데이터 값의 개수 B = 포착되어야 할 모든 데이터 값의 개수
이상 기준	데이터 손실로 인한 불완전성	센서 데이터 집합에 누락된 값이 있거나, 기대되는 실제 값 대신 0과 같은 기본값을 포함하고 있는 정도	X = A / B A = 누락되었거나 기대 값 대신 기본값인 데이터 값의 개수 B = 포착되어야 할 모든 데이터 값의 개수
	데이터 양 부족으로 인한 불완전성	특정 사용 맥락에서 포착된 데이터 값의 개수가 최소 요구 개수에 미치지 못하는 정도	X = max {0, 1 – A / B} A = 실제로 포착된 데이터 값의 개수 B = 포착되어야 할 최소 데이터 값의 개수

B.2.4 일관성을 위한 품질 측정지표

일관성 품질 측정지표는 센서 데이터가 시간에 따라 변하는 패턴이나 측정 센서가 생성하는 값에 적용되는 규칙을 얼마나 잘 만족하는지를 나타낸다(표 B-4참조). 이러한 규칙은 단일 센서의 데이터 패턴 뿐 아니라 여러 센서 간의 관계에서도 도출된다. 또한, 단일 센서에서 발생하는 편이, 범위 진동, 값 고정, 급락/급상승과 같은 데이터 이상, 그리고 다중 센서에서 발생하는 이질성 및 규칙 위반도 포함된다. 품질 특성 및 데이터 이상에 특화된 품질 측정지표는 표 B-4에 요약되어 있다.

표 B-4: 일관성 및 관련 데이터 이상에 대한 품질 측정지표

구분	품질 측정지표	설명	계산식
품질 특성 기준	데이터 세트의 일관성	센서 데이터 세트가 시간 변화 패턴이나 센서가 생성하는 값에 대한 규칙을 만족하는 정도	X = A / B A = 해당 규칙을 만족하는 측정 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	데이터 세트 간의 일관성[9]	서로 다른 센서가 생성한 두 데이터 세트가 두 세트 간의 관계 규칙을 만족하는 정도	X = A / B A = 해당 규칙을 만족하는 측정 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
이상 기준	편이에 의한 비일관성	센서 데이터의 추세 또는 변화율이 요구사항을 위반하는 정도	X = A / B A = 추세 또는 변화율 요구사항을 위반하는 측정 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	급락/급상승에 의한 비일관성	센서 데이터가 사전에 정의된 정상 패턴 이전 또는 이후에 급격히 감소하거나 증가하는 정도	X = A / B A = 사전 정의된 정상 패턴 범위를 벗어난 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	값고정에 의한 비일관성	데이터 값이 사전 정의된 정상 패턴보다 더 오랜 기간 동안 거의 변하지 않거나 전혀 변하지 않는 정도	X = A / B A = 정상 패턴보다 변화가 적고 지속 기간이 긴 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	범위 진동에 의한 비일관성	데이터 세트에서 한 시점과 다음 시점 간 값의 전체 범위가 변화하는 정도	X = A / B A = 사전 정의된 정상 범위를 벗어난 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	이질성에 의한 비일관성	서로 다른 센서의 두 데이터 세트가 동일해야 할 값에서 허용 가능한 차이를 넘어서는 정도	X = A / B A = 허용 가능한 차이를 초과한 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수
	규칙 위반에 의한 비일관성	서로 다른 센서의 두 데이터 세트가 사전에 정의된 관계나 규칙을 위반하는 정도	X = A / B A = 사전 정의된 규칙을 위반한 데이터 값의 개수 B = 평가된 전체 데이터 값의 개수

B.2.5 정밀성을 위한 품질 측정지표

1) 표현 정밀성

표현 정밀성 품질 측정지표는 센서 데이터 값이 정확하거나 특정 사용 맥락에서 충분히 구분할 수 있을 정도로 정밀한지를 나타낸다. 이는 해상도 차이와 같이 표현 정밀성에 영향을 미칠 수 있는 데이터 이상도 포함한다. 품질 특성 및 데이터 이상에 특화된 측정지표는 표 B-5에 요약되어 있다.

[9] 데이터 세트 간 일관성 측정지표의 경우, 지정된 측정 함수는 측정지표 설명을 충족한다. 그러나 이 함수는 일관성 또는 유사성을 측정하는 특정 방법에 대응하는 허용 기준(acceptance criteria)에 의존한다. 예를 들어, 유클리드 거리(Euclidean distance), 코사인 유사도(cosine similarity), 동적 시간 왜곡(dynamic time warping)과 같은 방법이 있다. 각 분석 방법은 서로 다른 유사성 값을 생성한다. 유클리ㄷ 거리 또는 동적 시간 왜곡 방법에서는 유사성이 0 이상이며, 값이 0일 때 두 데이터 세트가 가장 유사하고, 값이 클수록 유사성이 낮다. 코사인 유사도 방법에서는 유사성 값이 [0,1] 범위에 있으며, 값이 1일 때 두 데이터 세트가 가장 유사하고, 값이 0일 때 가장 불일치하다. 각 방법은 특정 도메인이나 맥락에 적합하며, 해당 방법에 맞는 허용 기준을 설정해야 한다.

표 B-5: 표현 정밀성 및 관련 데이터 이상에 대한 품질 측정지표

구분	품질 측정지표	설명	계산식
품질 특성 기준	데이터 세트의 표현 정밀성.[10]	데이터 값이 소수 자릿수(decimal places) 요구사항을 충족하는 정도	X = A / B A = 요구되는 표현 정밀성를 갖춘 측정 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
이상 기준	센서 해상도 불일치에 따른 부정밀성.[11]	데이터 값이 소수 자릿수 요구사항을 위반하는 정도	X = A / B A = 소수 자릿수 요구사항을 위반한 측정 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수

2) 측정 정밀성

측정 정밀성은 센서 데이터가 데이터 값 집합의 분산을 기반으로 한 무작위 분포의 지정된 신뢰 범위(confidence bound) 내에 있는 정도를 나타낸다. 이 품질 측정지표는 스파이크(spike)나 잡음(noise)과 같이 측정 정밀성에 영향을 줄 수 있는 데이터 이상도 포함한다. 관련 측정지표는 표 B-6에 요약되어 있다.

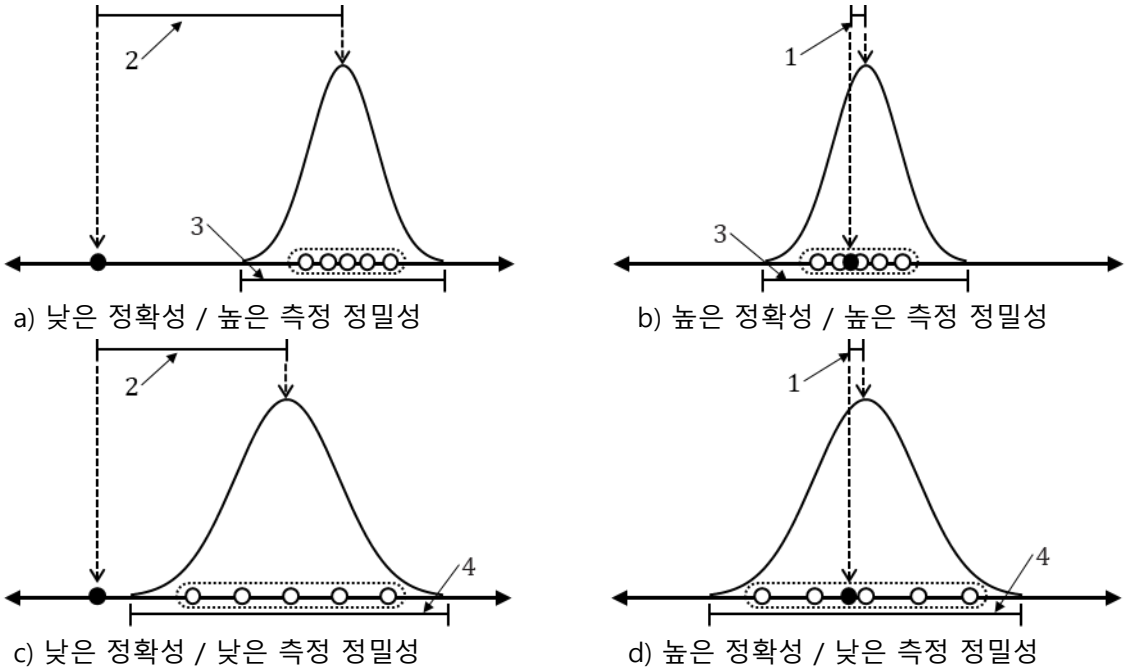
표 B-6: 측정 정밀성 및 관련 데이터 이상에 대한 품질 측정지표

구분	품질 측정지표	설명	계산식
품질 특성 기준	데이터 세트의 측정 정밀성.[12]	센서 데이터 세트가 무작위 분포의 지정된 신뢰 범위 내에 포함되는 값으로 이루어진 정도	X = A / B A = 무작위 분포의 지정된 신뢰 범위 내에 있는 측정 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
이상 기준	스파이크에 의한 부정밀성	센서 데이터 값이 정상 또는 사전에 정의된 패턴에 해당하는 신뢰 구간을 벗어나는 정도	X = A / B A = 신뢰 구간을 벗어난 측정 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
	노이즈에 의한 부정밀성	센서 데이터 값이 의도된 특성의 실제 표현에 해당하는 신뢰 구간을 무작위로 벗어나는 정도	X = A / B A = 해당 신뢰 구간을 무작위로 벗어난 측정 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수

[10] ISO/IEC 25024에서 발췌.

[11] 예: 유효 숫자가 5자리로 제한된 대기압 센서가 실제 값 10.1234, 53.1813, 89.6654, 100.1234, 178.6642, 210.9538을 포착하면, 데이터는 각각 10.123, 53.181, 89.665, 100.12, 178.66, 210.95로 기록된다. 이 데이터는 동일한 소수 자릿수를 갖고 있지 않다(100 미만의 값은 소수점 이하 3자리, 100 이상의 값은 소수점 이하 2자리만 유지). 대기압 값이 소수점 이하 3자리 해상도를 가져야 한다는 요구사항이 있는 경우, 소수 자릿수 부족으로 인한 부정밀성는 이 데이터 세트에서 50%가 된다.

[12] 품질 특성인 정확성(accuracy)과 측정 정밀성(measurement precision)는 서로 다르다(그림 B-1 참조). 포착된 데이터는 의도된 특성의 실제 표현과의 거리에 따라 낮거나 높은 정확성을 갖는다. 정확성에 대한 사전 요구사항이나 기준이 없으면, 데이터의 정확성은 상대적인 개념이다. 그러나 요구사항이나 기준이 있으면, 데이터는 명확하게 정확하거나 부정확하다. 이 구분은 측정 정밀성에도 동일하게 적용된다. 데이터의 변동 범위에 따라 낮거나 높은 정밀성을 갖지만, 사전 요구사항이나 기준(예: 허용 범위)이 존재할 경우, 데이터는 명확하게 정밀하거나 부정밀하다.



범례(Key)
○ 측정된 값 ● 실제 값
1: 높은 정확성 2: 낮은 정확성 3: 높은 측정 정밀성 4: 낮은 측정 정밀성

그림 B-9: 정확성과 측정 정밀성의 비교

B.2.6 적시성을 위한 품질 측정지표

적시성 품질 측정지표는 센서 데이터가 해당 속성 값이 발생한 시점을 정확히 나타내는 타임스탬프를 포함하는 정도를 나타낸다(표 B-7참조). 이러한 품질 측정지표는 지연(latency), 주기 불일치(inconsistent frequency), 잘못된 타임스탬프(incorrect timestamp), 타임스탬프 불일치(inconsistent timestamp) 등 적시성에 영향을 줄 수 있는 데이터 이상도 포함한다. 품질 특성 및 데이터 이상에 특화된 측정지표는 표 B-7에 요약되어 있다.

표 B-7: 적시성 및 관련 데이터 이상에 대한 품질 측정지표

구분	품질 측정지표	설명	계산식
품질 특성 기준	데이터 세트의 적시성	데이터 값이 허용 가능한 시간 한계 내의 타임스탬프를 갖는 정도	X = A / B A = 허용 가능한 시간 한계 내의 타임스탬프를 가진 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
	데이터 세트 간의 적시성	동시에 수집된 값들의 타임스탬프가 서로 동일한 정도	X = A / B A = 두 데이터 값의 타임스탬프가 동일한 경우의 개수 B = 동일해야 할 것으로 평가된 타임스탬프의 총 개수
이상 기준	지연에 의한 비적시성	센서 데이터 값의 타임스탬프가 기대보다 지속적으로 늦는 정도	X = A / B A = 기대보다 늦은 타임스탬프를 가진 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
	수집 주기 불일치에 의한 비적시성	데이터 수집 주기가 사전 정의된 주기에서 변경되거나 벗어나는 정도	X = A / B A = 사전 정의된 수집 주기를 위반하는 타임스탬프를 가진 데이터 값의 개수

			B = 측정된 전체 데이터 값의 개수
	잘못된 시간정보에 의한 비적시성[13]	시간정보가 잘못되었거나 유효하지 않은 정도	X = A / B A = 형식이 잘못되었거나 잘못된 값을 가진 타임스탬프의 데이터 값 개수 B = 측정된 전체 데이터 값의 개수
	순서 오류 (잘못된 시간정보)에 의한 비적시성	타임스탬프가 올바른 순서를 따르지 않는 정도	X = A / B A = 올바른 순서에서 벗어난 타임스탬프를 가진 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
	시간정보 편차 (잘못된 시간정보)에 의한 비적시성	타임스탬프가 기대되는 값에서 벗어나는 정도	X = A / B A = 기대값에서 벗어난 타임스탬프를 가진 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
	시계 드리프트 (잘못된 시간정보)에 의한 비적시성	기준 시계와 센서 노드 또는 시스템 간의 시차가 존재하는 정도	X = A / B A = 기준 시계와 센서 노드/시스템 간의 시간 차이를 가진 데이터 값의 개수 B = 측정된 전체 데이터 값의 개수
	시간정보 불일치에 의한 비적시성	동시에 수집된 두 데이터 값의 타임스탬프가 서로 다른 정도	X = A / B A = 동시에 수집된 값에서 시간 정보가 동일하지 않은 데이터 값의 개수 B = 동시에 수집된 데이터 값의 총 개수
	시간대 불일치 (잘못된 시간정보)에 의한 비적시성	시간정보의 시간대 정보 문제로 인해 센서 데이터가 올바르게 해석될 수 없는 정도	X = A / B A = 시간대 문제를 가진 시간정보 데이터 값 개수 B = 측정된 전체 데이터 값의 개수

[13] 단일 센서가 동일한 타임스탬프를 가진 두 개 이상의 데이터 값을 수집하는 경우, 이는 잘못된 타임스탬프의 사례이다.

C. 이상 데이터의 정제 기법

C.1 이상 데이터의 탐지

C.1.1 데이터 이상 및 탐지 유형

데이터 이상은 세 가지 경우로 분류될 수 있다:

- 점 이상: 특정 데이터 인스턴스가 나머지 데이터와 비교해 이상으로 간주될 수 있다면, 해당 인스턴스는 점 이상으로 정의된다.
- 집합적 이상: 관련 데이터 인스턴스의 집합이 전체 데이터 세트나 패턴에 비해 이상으로 판단될 경우, 이를 집합적 이상이라고 한다.
- 맥락적 이상: 특정 맥락(context)에서 데이터 인스턴스가 이상으로 분류되지만 다른 맥락에서는 그렇지 않은 경우, 이를 맥락적 이상(또는 조건적 이상)이라고 한다.

이상 탐지 접근 방식은 과거 역사적 데이터로부터의 모델과 예측에 기반을 두고 있다. 이상 탐지 알고리즘을 적용할 때 세 가지 가능한 경우가 고려될 수 있다:

- 정확한 탐지: 탐지된 데이터 이상은 실제 현장에서 발생한 이상과 정확히 일치한다.
- 거짓 양성 (false positives): 실제 필드는 계속해서 정상이지만 시스템 오류 및 오작동 등으로 인해 예상치 않게 비정상적인 데이터 값으로 나타난다.
- 거짓 부정 (false negatives): 실제 필드는 비정상이지만 시스템 오류 및 오작동 등으로 인해 예상치 않게 정상적인 데이터 값으로 나타난다.

이 문서에서는 데이터 정제가 유지된 센서 데이터에서 데이터 이상이 탐지될 때만 가능하기 때문에, 정확한 탐지와 거짓 양성이 고려될 것이다.

C.1.2 시계열 데이터의 이상 탐지[14]

시계열은 시간 스탬프와 연관된 데이터 항목(수치 값)이 완전히 순서화 된 데이터이다. 이 시간 스탬프는 두 항목 사이의 시간 간격을 식별할 수 있도록 한다. 따라서 센서 데이터는 단일하고 이산적인 디지털 값의 스트림으로, 시계열의 한 유형이다.

시계열 데이터에 대한 이상 탐지 방법은 다양하며, 여기서는 이 중 21가지 잘 알려진 방법이 제시되어 있다. 이들은 기본(basic), 통계적(statistical), 디지털 신호 처리(digital signal processing), 기계 학습(machine learning)의 네 가지 범주로 분류되었다.

C.1.2.1 기본 방법 유형

여기에는 고정 임계값과 동적 임계값과 같은 여러 가지 방법이 포함된다. 각 방법은 아래에서 설명된다:

- 고정 임계값 (fixed threshold)

[14] 참고: 이 이상 탐지 방법은 이상 데이터 값이나 패턴을 이상으로 탐지하기 위해 설계되었으며, 데이터 이상 유형을 식별하기 위한 것은 아니다.

미리 정해진 정적 값인 임계값을 사용하여 이상을 식별하는 기술이다. 도메인 지식, 역사적 데이터 분석 또는 기타 관련 기준을 기반으로 하한값과 상한값이 임계값으로 설정된다. 시계열의 새로운 데이터 포인트가 측정되면 이 고정 임계값과 비교된다. 측정된 값이 상한 임계값을 초과하거나 하한 임계값 아래로 떨어지면 이상으로 표시된다.

– 동적 임계값 (dynamic threshold)

동적으로 조정되는 임계값을 사용하여 이상값을 식별하는 방법이다. 임계값은 신호의 통계적 특성(평균 및 분산 등)과 현재 또는 최근 기간의 배경 노이즈 수준에 따라 적응적으로 조정된다. 데이터 포인트가 이 동적 임계값보다 크거나 작을 경우 이상값으로 간주된다 [11].

– 시간 간격 분석 (time interval analysis): 시간 표시 데이터 세트에서 연속된 시간 표시 데이터 간의 간격을 기초로 이상값을 식별하는 방법이다. 이 방법은 인접한 시간 표시 데이터 간의 차이를 계산하고 이러한 간격을 사전 정의된 임계값과 비교한다. 임계값을 초과하는 간격은 이상값으로 표시되며, 이는 잘못된 시간 표시나 데이터 손실을 의미할 수 있다.

– 순차적 의존성 검사 (sequential dependency check): 타임스탬프의 정확성을 검증하기 위해 타임스탬프가 사건의 chronological(시간적) 및 logical(논리적) 순서와 일치하는지 확인하는 방법이다. 이 방법은 타임스탬프의 시간적 순서와 관련된 사건의 논리적 순서를 분석하여 누락된 데이터, 불필요한 데이터, 또는 순서가 뒤바뀐 데이터를 식별한다.

– 슬라이딩 윈도우 (sliding window): 입력 데이터에서 윈도우 또는 범위를 정의한 후 해당 윈도우를 데이터 전체에 걸쳐 이동시켜 윈도우 내부의 특정 작업을 수행하는 기본적인 방법이다. 이 방법은 데이터 세트의 끝까지 윈도우를 한 요소씩 오른쪽으로 이동시킨다. 각 윈도우에 대해 평균과 표준 편차를 계산한 후 데이터 포인트를 임계값(예: $\mu \pm 3\sigma$)과 비교한다. 임계값보다 크거나 작은 데이터 포인트는 이상값으로 간주된다.

C.1.2.2 통계적 방법 유형

여기에는 주성분 분석과 사분위수 범위 등 여러 가지 방법이 포함된다. 각 방법은 아래에서 설명된다:

– 주성분 분석 (principal component analysis): 선형 차원 축소 기법 중 하나로, 데이터 세트를 주성분이라고 불리는 새로운 특징 집합으로 변환하는 기술이다. 차원 축소 기법을 사용하면 원본 데이터에서 주요 성분이 추출되며, 이후 이 주요 성분 중 일부만을 사용하여 원본 데이터를 재구성한다. 재구성 과정에서 재구성 오류가 큰 데이터 항목은 이상값으로 간주된다 [12].

– 사분위수 범위 (inter-quartile range): 사분위수 범위(IQR)를 사용하여 이상값을 탐지하는 통계적 기술이다. IQR은 데이터의 분산 정도를 측정하는 지표로, 데이터의 분포를 의미하며, 데이터의 75번째와 25번째 백분위수 사이의 차이로 정의된다. 25번째 백분위수는 첫 번째 사분위수(Q1)로, 75번째 백분위수는 세 번째 사분위수(Q3)로도 알려져 있다. IQR을 계산하려면 데이터 세트를 사분위수로 나누어 데이터 포인트의 수를 대략적으로 동일한 크기의 네 부분으로 나누게 된다. 측정된 값이 Q1에서 IQR의 1.5배를 뺀 값보다 작거나 Q3에 IQR의 1.5배를 더한 값보다 큰 경우 해당 값은 이상값으로 탐지된다 [13].

– 지역 이상값 요인 (Local outlier factor, LOF): 주어진 데이터 포인트의 이웃 데이터 포인트에 대한 상대적 밀도를 기반으로 이상값을 탐지하는 비지도 학습 기반 이상값 탐지 기법이다. 로컬 밀도는 일반적으로 데이터 포인트에서 k-최근접 이웃까지의 평균 거리의 역수로 정의된다. 데이터 포인

트의 로컬 아웃라이어 팩터(LOF) 점수는 해당 데이터 포인트의 로컬 밀도와 그 이웃들의 로컬 밀도를 비교하여 계산된다. 데이터 포인트의 LOF 점수가 1보다 현저히 높다면, 이는 해당 데이터 포인트가 이웃들에 비해 현저히 낮은 로컬 밀도를 가지고 있음을 나타내며, 이 경우 해당 데이터 포인트는 이상값으로 간주된다 [14].

- 표준 점수 판정 기법 (z-score based technique): z-점수는 표준 점수라 불리며, 데이터 포인트가 데이터 집합의 평균으로부터 표준 편차의 몇 배 만큼 떨어져 있는지를 나타내는 수치이다. z-점수가 임계값을 초과할 경우 이상값으로 간주된다. 일반적으로 z-점수가 3보다 크거나 -3보다 작은 경우를 임계값으로 사용하며, 이는 해당 데이터 점이 평균으로부터 3 표준편차 이상 떨어져 있음을 의미한다 [15].
- 삼 요소 분해 기법 (tri-class anomaly detection, Tri-CAD): 시계열 자료를 추세, 계절성, 잔차 성분의 요소로 분리한 후, 각 요소 별로 이상값을 식별하는 기술. 이 분해 방법은 로컬 가중 산점도 평활화(LOESS)를 통해 구현되며, 이는 데이터의 다양한 구간에서 다중 회귀 분석을 적용하여 매끄러운 곡선을 생성하는 비모수적 회귀 방법이다. 이상 탐지 시에는 주로 잔차 성분에 초점을 맞춘다. 잔차의 크기에 임계값을 설정하여 이상을 식별할 수 있다. 잔차가 예상 범위(예: 잔차의 평균으로부터 특정 표준편차 이상)에서 크게 벗어난 데이터 포인트는 잠재적 이상으로 표시된다 [16].
- 1차 자기회귀 모델 (first-order autoregressive): 시계열 자료의 각 관측값을 이전 관측값과 일부 임의의 잡음의 선형 함수로 모델링하는 방법이다. 이상 탐지 분야에서 1차 자기회귀 모델은 마지막 관측값을 기반으로 시계열의 다음 값을 예측하는 데 사용된다. 예측 오차(잔차)는 실제 관측값과 예측값의 차이로 정의된다. 잔차의 절대값 또는 제곱 잔차가 사전 정의된 임계값을 초과할 때 이상이 탐지된다 [17].
- 베이즈 기반 변화점 탐지 (Bayesian-based change point detection): 베이즈 통계 이론을 사용하여 시계열의 기본 분포에서 급격한 변화나 변동을 식별하는 접근법이다. 시계열에서 이상 데이터 탐지에 효과적인 방법이다. 모델을 합리적으로 정의하고, 사후 분포를 업데이트하고, 변화점의 확률을 계산하고, 이상 데이터를 식별함으로써 시계열 데이터 내 이상 데이터를 정확하게 탐지할 수 있다 [18].

C.1.2.3 디지털 신호 처리 방법 유형

여기에는 적응형 입자 필터링, 웨이블릿 변환 등 여러 가지 방법이 포함된다. 각 방법은 다음과 같다:

- 적응형 입자 필터링 (adaptive particle filtering): 입자 필터링의 확장된 형태로, 입자 그룹을 사용하여 시스템 상태 분포를 표현하고, 이 과정을 향상시키기 위해 입자의 가중치를 동적으로 조정한다. 이 기법은 상태 공간 모델의 사후 분포를 더욱 정확하게 추정할 수 있다. 센서 출력을 시스템 상태에 연결하는 모델을 구축하고, 센서 출력을 기반으로 입자 가중치를 업데이트하며, 허용 가능한 편차 임계값을 설정하고, 실제 센서 출력과 모델 예측값의 차이가 임계값을 초과하는 데이터는 플래그를 지정하여 잠재적 이상 징후를 표시한다 [19].
- 웨이블릿 변환 (wavelet transform): 신호를 여러 시간-주파수 성분으로 분해하여 다양한 스케일의 특징을 추출하는 수학적 도구이다. 센서 신호의 다중 스케일 분해를 통해 다양한 시간 스케일의 특징을 추출하고, 정상 작동 중 수집된 데이터를 기반으로 임계값 모델을 구축한다. 특정 스케일에서 측정된 값의 계수가 미리 설정된 임계값을 초과하면 신호는 비정상으로 간주된다 [20].

– 푸리에 변환 (Fourier transform): 푸리에 변환을 사용하여 시간 영역에서 주파수 영역 표현으로 신호를 변환하는 수학적 변환이다. 시계열을 구성 주파수로 분해하여 원래 시간 영역 표현에서는 나타나지 않는 주기성, 주요 주파수 및 기타 스펙트럼 특성을 식별할 수 있다. 이상 탐지의 맥락에서 푸리에 변환은 주파수 영역에서 예상치 못하거나 비정상적인 패턴을 강조하여 이상 신호를 드러낼 수 있다 [21].

C.1.2.4 기계학습 방법 유형

여기에는 확률 신경망, 자기회귀 통합 이동 평균 모델 등 여러 가지 방법이 포함된다. 각 방법은 다음과 같다:

– 확률적 신경망 (probabilistic neural network): 분류 및 패턴 인식 문제에 널리 사용되는 피드포워드 신경망이다. 여러 확률 밀도 함수를 구성하여 다양한 패턴의 학습 샘플을 학습한다. 이 방법은 센서 입력과 학습된 정상 패턴 간의 일치 확률을 계산하여 이상 징후를 감지할 수 있으며, 일치 확률이 낮은 이상 징후는 이상 징후로 간주된다 [22].

– 시계열 분석에 사용되는 통계 모델: 자기회귀, 차분, 이동평균의 세 가지 구성 요소를 결합한다. 자기회귀 부분은 과거 관측치를 기반으로 미래 값을 예측한다. 차분은 추세를 제거하여 시계열을 정상 상태로 만드는 데 사용된다. 이동평균 부분은 예측의 오차항 또는 잔차를 모델링한다. 이상 탐지에서는 자기회귀 통합 이동평균 모델을 사용하여 시계열 데이터를 예측할 수 있다. 예측된 값을 실제 관측치와 비교하여 유의미하게 벗어난 지점을 식별하는 데 도움이 된다. 이러한 편차는 이상값으로 간주된다 [23].

– 격리 숲 (isolation forest): 비지도 학습 이상 탐지 알고리즘으로, 특징을 무작위로 선택하고 데이터 포인트를 분할하여 이상값을 격리한다. 이 알고리즘은 무작위 데이터 하위 집합을 사용하여 여러 개의 의사결정 트리를 구축한다. 이상값 점수는 데이터 포인트를 격리하는 데 필요한 평균 분할 횟수를 기반으로 한다. 분할 횟수가 적을수록 이상값일 가능성이 높다 [24].

– 단일 클래스 지원 벡터 머신 (one-class support vector machine): 이상 탐지를 위해 설계된 비지도 학습 알고리즘이다. 단일 클래스의 데이터 만을 사용하여 학습하며, 정상 데이터 포인트를 포함하는 고차원 공간에서 최적의 초구(hypersphere)를 찾는 것을 목표로 한다. 이 초구 밖에 있는 모든 테스트 데이터 포인트는 이상(anomaly)으로 분류된다 [25].

– 오토인코더 (autoencoder): 입력 데이터(일반적으로 레이블이 지정되지 않은)의 효율적인 코딩을 학습하도록 설계된 신경망 아키텍처로, 인코더와 디코더라는 두 가지 주요 구성 요소로 구성된다. 시계열 이상 탐지를 위해 오토인코더는 정상(비이상) 데이터를 기반으로 학습되어 일반적인 패턴을 효과적으로 재구성하는 방법을 학습한다. 정상 데이터는 오토인코더가 이러한 패턴을 재구성하는 방법을 이미 학습했기 때문에 재구성 오류가 작다. 따라서 재구성 오류에 임계값을 설정하여 이 임계값을 초과하는 오류를 가진 인스턴스를 잠재적 이상(anomaly)으로 표시할 수 있다 [26]

– 순환 신경망 (recurrent neural network, RNN): 노드 간 연결이 시간적 시퀀스를 따라 유향 그래프를 형성하는 인공 신경망이다. 피드포워드 신경망에서 파생된 순환 신경망(RNN)은 내부 상태(메모리)를 사용하여 가변 길이 입력 시퀀스를 처리할 수 있다. RNN은 이상값 없이 과거 데이터에 적합하여 시계열의 정상적인 동작을 학습한다. 모델은 이전 값을 기반으로 시퀀스의 다음 값을 예측하는 방법을 학습한다. 이상을 감지하기 위해 예측값과 실제 측정값의 차이를 계산한다. 이 차이

는 절대값 차이 또는 제곱값일 수 있다. 이 차이가 미리 정의된 임계값을 초과하면 데이터 포인트는 이상으로 표시된다 [27].

C.1.3 이상 탐지 방법으로 탐지 가능한 이상 유형

C.1.2에서 설명한 바와 같이, 각 이상 탐지 방법은 하나 이상의 이상 유형을 처리할 수 있다. 각 이상 탐지 방법이 탐지할 수 있는 이상 유형은 표 C-1에 요약되어 있다.

표 C-1: 이상 탐지 방법으로 탐지 가능한 이상 유형

No	Type of **Detection Method**	**Detection Method**	**Detectable Anomaly Type**
1	Basic	Fixed threshold	Offset
2		Dynamic threshold	Offset
3		Time interval analysis	Incorrect timestamps, Data loss
4		Sequential dependency check	Incorrect timestamps
5		Sliding window	Incorrect timestamps, Inconsistent frequency
6	Statistical	Principal component analysis	Drift
7		Inter-quartile range	Spike
8		Local outlier factor	Spike
9		Z-score	Spike
10		Seasonal and trend decomposition using Loess	Drop or rise, Spike
11		First order auto regressive	Shift
12		Bayesian-based change point detection	Shift
13	Digital signal processing	Adaptive particle filtering	Spike, Trim
14		Wavelet transform	Spike, Trim
15		Fourier transform	Noise, Inconsistent frequency
16	Machine learning	Probability neural network	Drift
17		Auto regressive integrated moving average model	Shift, Spike, Trim
18		Isolation forest	Spike
19		One-Class Support vector machine	Drop or rise, Spike
20		Autoencoder	Shift
21		Recurrent neural network	Drop or rise

C.2 데이터 이상 보정

C.2.1 이상 데이터의 보정 유형

데이터 이상 징후가 탐지되면 이상 징후 데이터의 보정이 추가로 고려된다. C.1.1에서 설명한 바와 같이, 탐지된 데이터 이상 징후는 정상 탐지 또는 거짓양성(false positive)이다. 거짓양성의 경우, 시스템 장애 또는 오작동으로 인해 예상치 못한 이상 징후 데이터 값이 관찰되므로, 이상 징후 데이터를 삭제하거나 적절한 데이터 값으로 대체하여 보정할 수 있다. 정상 탐지된 경우에도 데이터 사용 목적에 따라 이상 징후 데이터를 보정할 수 있다. 이상 징후 데이터를 보정할 때는 원래 정상 데이터가 변경되지 않도록 이해관계자의 동의를 얻는 것이 좋다.

C.2.2 시계열 데이터 이상 보정

이 섹션에서는 시계열 데이터의 이상을 수정하는 30가지 기존 방법을 제시하며, 이는 기본, 통계, 디지털 신호 처리, 머신 러닝의 네 가지 범주로 분류된다.

C.2.2.1 기본 방법 유형

여기에는 참조 데이터 비교 및 보간 대체와 같은 여러 가지 방법이 포함된다. 각 방법은 다음과 같다:

- 참조 데이터 비교 (referenced data comparison): 과거 데이터 또는 알려진 표준 값을 기준 값으로 선택하는 기준선 방식이다. 센서 데이터의 각 값과 기준 값의 차이를 계산하고, 이 차이를 이상 데이터에 더하여 드리프트 및 진동을 보정한다.
- 평균/중앙값/모드 대체 (mean/median/mode imputation): 결측값을 채우는 데 사용되는 가장 쉬운 방법 중 하나이다. 이는 각 결측값을 동일한 속성에 대한 관측 데이터의 평균, 중앙값 또는 모드값으로 대체하여 수행된다 [28].
- 보간 대체 (interpolation replacement): 보간법을 사용하여 데이터 집합의 이상값을 추정하고 대체하는 방법이다. 보간법은 알려진 데이터 집합을 기반으로 새로운 데이터 포인트를 구성하는 기법이다. 측정된 데이터 포인트가 선형 또는 다항식 관계와 같은 특정 함수 관계를 따른다고 가정하며, 따라서 선형 함수, 다항식 함수 등을 사용하여 적합할 수 있다. 적합된 곡선에서 데이터 포인트를 추출하여 데이터가 누락된 부분을 채우거나 시퀀스에서 이상 데이터 포인트를 대체한다.
- 리샘플링 (resampling): 리샘플링을 통해 신호 샘플링 주파수를 조정하여 이상값을 수정하는 기법이다. 비일관적인 이상값을 수정하기 위해 목표 샘플링 주파수를 설정하고, 리샘플링을 통해 원래 신호를 이 목표 주파수로 변환한다. 결과적으로, 비일관적인 주파수를 가진 데이터를 수정하여 전체 데이터 세트에서 균일한 샘플링 주파수를 얻을 수 있다.
- 동기화 (time synchronisation): 신뢰할 수 있는 정확한 시간 데이터 소스와 타임스탬프를 일치시켜 잘못된 타임스탬프를 수정하는 방법이다. 정확한 타임스탬프를 가진 것으로 알려진 다른 데이터 소스와 타임스탬프를 비교하여 시간 오프셋을 추정하고, 타임스탬프를 조정하여 잘못된 타임스탬프를 수정한다.
- 타임스탬프 보간 (timestamp interpolation): 데이터 집합 내에서 누락된 타임스탬프를 추정하고 삽입하여 완전하고 연속적인 시계열을 보장하는 방식이다. 누락된 타임스탬프는 주변 타임스탬프를 기반으로 추정 및 삽입되어 완전한 데이터 시계열을 복원한다.

C.2.2.2 통계 방법 유형

여기에는 최소제곱법, 다항식 피팅 등 여러 가지 방법이 포함된다. 각 방법은 다음과 같다:

- 최소제곱법 (least squares): 관측값과 모델 예측값 사이의 오차 제곱의 합을 최소화하여 최적의 적합 함수를 찾는 수학적 최적화 기법이다. 센서에서 수집된 시계열 데이터를 다룰 때 데이터 누락, 노이즈, 드리프트와 같은 문제가 발생할 경우, 최소제곱법을 사용하여 적절한 모델을 추정하고, 이 모델을 기반으로 데이터의 이상값을 수정할 수 있다 [29].
- 다항식 피팅 (polynomial fitting): 특정 차수의 다항식 함수를 사용하여 데이터 포인트 집합을 근사하는 데 사용되는 통계 기법이다. 다항식 피팅을 수행하려면 일반적으로 시간 변수와 관측값 간의 관계를 가장 잘 설명하는 특정 차수의 다항식을 선택한다. 이 과정은 실제 데이터 포인트와 다항식 함수로 예측된 값의 차이 제곱의 합을 최소화하는 다항식의 계수를 찾는 과정을 포함한다. 다항식을 데이터에 피팅한 후, 피팅된 곡선의 경계를 크게 벗어나는 모든 데이터 포인트는 이상값으로 간주할 수 있다. 이상값을 식별한 후에는 예상 추세를 따르는 값으로 대체하여 수정할 수 있다.
- 지수 피팅 (exponential fitting): 지수 함수를 사용하여 데이터 포인트 집합을 근사하거나 피팅하는 데이터 분석 및 곡선 피팅 방법이다. 이 방법은 시계열 데이터의 일반적인 움직임을 포착하는 지수 모델을 구축하고, 최적화 기법을 사용하여 지수 피팅 함수의 매개변수를 추정하여 모델이 과거 데이터에 최대한 근접하도록 한다. 이 모델은 시계열의 기대값을 예측하는 데 사용될 수 있다. 이상값을 감지한 후, 이 방법은 이상값 데이터 포인트를 피팅된 모델의 예측값과 같은 더 합리적인 값으로 대체하여 이상값을 수정하는 것을 목표로 한다.
- 뎀스터-세이퍼 기법 (Dempster-Shafer): 기본 확률 할당을 사용하여 가설에 대한 지원 정도를 표현하는 결합 기술로, 신념 함수와 가능성 함수를 통해 증거의 품질을 측정한다. 동시에 지원 정도와 신뢰도에 따라 결합 규칙이 형성된다. 이 기술은 다중 센서에서 수집된 데이터를 평가하고 결합하여 이상값을 보정할 수 있다 [30].

C.2.2.3 디지털 신호 처리 유형

여기에는 칼만 필터링과 저역 통과 필터링 등 여러 가지 방법이 포함된다. 각 방법은 아래에서 설명된다:

- 칼만 필터링 (Kalman filtering): 시간에 따라 관측된 측정값의 시열을 활용해 알려지지 않은 변수의 추정치를 산출하는 대표적인 방법이다. 이 방법은 적응형 예측 알고리즘을 통해 재귀적 계산으로 과거 추정값과 새로운 측정 정보를 결합함으로써 데이터 추정치를 동적으로 최적화한다. 이 과정은 센서 상태를 지속적으로 업데이트하며, 최종적으로 이상을 예측값으로 대체한다 [31].
- 저역 통과 필터링 (low-pass filtering): 저주파 신호를 통과시키면서 미리 설정된 임계값을 초과하는 고주파 신호를 차단하거나 약화시키는 필터링 방법이다. 이 방법은 고주파 노이즈를 필터링하기 위해 주파수 임계값을 설정하며, 남은 신호는 주로 저주파 추세 성분을 포함하게 된다. 식별된 고주파 노이즈 데이터 포인트는 필터링된 신호에서 얻은 값이나 정상 데이터 포인트를 사용하여 수행된 보간법으로 대체된다.

- 대역통과 필터링 (band-pass filtering): 특정 주파수 범위 내의 주파수만을 선택적으로 통과시키고 해당 범위 외의 주파수를 감쇠시키는 필터링 방법이다. 노이즈가 포함된 이상값을 수정하기 위해 특정 주파수 범위 내의 측정 값을 통과시키고 다른 주파수 범위 내의 측정 값을 감쇠시키거나 차단한다. 주파수 범위를 설정함으로써 해당 범위 외의 노이즈 데이터가 필터링된다.
- 고역통과 필터링 (high-pass filtering): 신호의 저주파 성분을 감쇠하거나 차단하면서 고주파 성분을 통과시키는 필터링 방법이다. 이 방법은 센서에서 측정된 고주파 값을 통과시키면서 저주파 측정 값을 감쇠하거나 차단하며, 최종적으로 천천히 변하는 저주파 드리프트 성분을 필터링하여 제거하고 데이터 내의 고주파 성분을 유지하여 이상 값을 복원한다 [32].
- 중앙값 필터링 (median filtering): 순서 통계 이론을 기반으로 한 비선형 신호 처리 기술이다. 이 기술은 신호의 각 요소를 차례로 이동하며, 각 요소를 인접한 요소들의 중앙값으로 대체하는 방식으로 작동한다. 이 방법은 극단적인 값이나 이상값이 존재하는 스파이크 또는 "소금과 후추" 노이즈를 제거하는 데 특히 효과적이다 [33].
- 가중 이동 창 필터링 (weighted moving window filtering): 특정 크기의 창이 신호 위를 이동하며, 창 내의 각 데이터 포인트는 해당 가중치와 곱해진다. 이 가중치 값들의 합은 창 중앙의 값을 대체하는 데 사용된다. 가중치는 특정 기준에 따라 선택될 수 있으며, 예를 들어 중앙 값이나 최근 데이터에 더 큰 중요성을 부여하는 방식이 있다. 이 방법은 중요한 신호 특성을 유지하면서 노이즈가 많은 데이터를 부드럽게 하는 데 유용하다 [32].
- 적응형 필터링 (adaptive filtering): 입력 신호의 변화하는 특성에 따라 필터 파라미터가 자동으로 조정되는 동적 신호 처리 방법이다. 적응형 필터는 실시간으로 동작을 조정하여 최적의 성능을 달성할 수 있다. 이 조정은 일반적으로 오류 신호에 의해 안내되며, 최소 평균 제곱법과 같은 알고리즘이 필터 계수를 업데이트하는 데 널리 사용된다. 적응형 필터링 방법은 노이즈 제거, 스파이크 제거, 신호 강화 등과 같은 응용 분야에서 널리 사용된다 [34].
- 웨이블릿 변환 (wavelet transform): 신호를 분해하여 주파수 및 시간 정보를 표시하는 신호 분석 기법이다. 이 기법은 시간에 따라 변하는 신호를 분석하는 데 특히 유용하며, 다양한 스케일에서 세부적인 분석을 가능하게 한다. 신호에서 변화나 특징이 발생하는 위치와 시점을 국소적으로 파악하여 스파이크 데이터를 나타내는 고주파 성분을 효과적으로 제거할 수 있다 [35].
- 경험적 모드 분해 (empirical mode decomposition): 시계열을 유한한 수의 성분으로 분해하는 방법으로, 각 성분은 데이터의 고유한 진동 모드를 나타낸다. 복소수 시계열을 여러 고유모드 함수와 최종 추세항으로 분해한다. 각 고유모드 함수는 측정값의 주파수 성분을 나타내며, 추세항은 일반적으로 저주파 성분을 포함한다. 이러한 분해 후, 신호의 추세항을 제거하여 신호의 단기 변동이나 주기적 특징을 분리할 수 있다. 이 방법은 센서 데이터의 드리프트, 노이즈 및 기타 이상 현상을 효과적으로 보정할 수 있다 [36].
- 압축 센싱 (compressed sensing): 미결정 선형 시스템의 해를 찾아 신호를 효과적으로 수집하고 재구성하는 신호 처리 기술이다. 이는 신호의 희소성을 활용하여 나이퀴스트-섀넌 샘플링 정리에 필요한 것보다 훨씬 적은 샘플에서 신호를 보정할 수 있다는 원리에 기반한다. 원래 측정값을 재구성하여 센서에 고정된 측정값을 보정할 수 있다 [37].

– 푸리에 변환 (Fourier transform): 원본 시계열 데이터를 주파수 영역으로 변환하는 방법으로, 데이터는 서로 다른 진폭과 위상을 갖는 사인 함수의 합으로 표현된다. 이상값은 주파수 성분을 분석하고 수정하여 수정한 후, 역푸리에 변환을 수행하여 이상값이 수정된 시계열을 재구성한다 [38].
– 동적 시간 왜곡을 기반으로 한 시간 정렬 (time alignment based on dynamic time warping): 두 시간 시리즈 간의 일관되지 않은 타임스탬프를 수정하기 위해 두 시간 시리즈 간의 최적 정렬을 찾는 접근 방식이다. 두 시간 시리즈의 대응되는 데이터 포인트 간의 거리를 계산하고 누적 거리를 최소화하여 최적의 왜곡 경로를 구한 후, 이 방법은 최적의 왜곡 경로를 기반으로 한 비균일 시간 왜곡을 사용하여 두 시퀀스를 정렬한다 [39].

C.2.2.4 기계학습 방법 유형

여기에는 이웃 기반 방법 및 회귀 기반 방법 등 여러 가지 방법이 포함된다. 각 방법은 다음과 같다:

– 이웃 기반 접근 (neighbour-based): 이 방법은 데이터 포인트의 로컬 이웃을 조사하여 이상 징후를 식별하고 수정하는 것을 추구한다. 이 방법의 주요 근거는 데이터 포인트(이웃)가 유사한 특성을 공유하므로 이 패턴에서 유의미한 편차가 발생하면 이상 징후로 볼 수 있다는 점이다. 이 경우 로컬 이웃 데이터 포인트의 평균값으로 대체될 수 있다고 가정한다. 일반적으로 사용되는 알고리즘으로는 k-최근접 이웃(k-NN)과 잡음이 있는 애플리케이션의 밀도 기반 공간 클러스터링(DBSCAN)이 있다 [40].
– 제약 기반 방법 (constraint-based): 이 방법은 시간 경과에 따른 데이터의 예상 동작을 반영하는 논리적 또는 통계적 제약 조건을 적용하여 시계열 데이터의 타임스탬프에서 불일치나 불규칙성을 감지하고 수정하는 방식을 취한다. 이러한 제약 조건은 데이터 자체에서 도출하거나 도메인 지식을 기반으로 정의할 수 있다. 타임스탬프가 예상 패턴을 준수하도록 함으로써 이 방법은 시계열 데이터의 일관성과 사용성을 유지하는 데 도움이 된다 [40] .
– 회귀 기반 방법 (regression-based): 이 방법은 사용 가능한 과거 데이터를 학습하여 결측값이나 이상값을 예측하는 회귀 모델을 학습하는 것이다. 이 방법은 현재 값과 과거 값 또는 데이터 세트의 다른 속성 사이에 관계가 있다고 가정한다. 시계열 데이터에 적용할 경우, 이 방법은 크게 간단한 자기회귀 모델과 더 진보된 반복적 접근 방식으로 분류될 수 있다. 회귀 기반 방법은 과거 데이터가 풍부하고 시계열의 무결성을 유지하기 위해 이상값을 정확하게 예측하고 수정해야 하는 상황에서 특히 유용하다[43].
– 행렬 분해 기반 방법 (matrix factorization-based): 이 수학적 방법은 원본 데이터 행렬을 두 개 이상의 저차원 행렬의 곱으로 분해하는 접근을 취한다. 이러한 분해는 데이터 내의 잠재적 관계와 상관관계를 밝히는 데 도움이 된다. 이러한 저차원 요인들로부터 원본 행렬을 재구성함으로써, 이 방법은 결측치 또는 이상값을 추정하여 시계열을 보정할 수 있다. 행렬 분해 기반 방법은 특히 희소 데이터를 처리하거나 변수 간의 근본적인 관계가 복잡하지만 저차원 잠재 요인으로 근사할 수 있는 경우에 유용하다 [40] .
– 기대 극대화 기반 방법 (expectation-maximization-based method): 이 방법은 반복적 EM(기대 극대화) 알고리즘을 활용하여 시계열 데이터 내의 이상값을 추정하고 수정하는 접근을 취한다. 이 방법은 관측 변수와 관측되지 않은(잠재) 변수를 모두 고려하여 시계열의 확률적 모델을 구성하는 것을 포함한다. EM 알고리즘은 현재 모델 매개변수를 고려하여 누락되거나 이상값을 추정하는 기

대(E) 단계와 모델의 가능도를 최대화하도록 모델 매개변수를 업데이트하는 최대화(M) 단계를 번갈아 수행한다. 이 반복적 프로세스는 추정치가 수렴할 때까지 계속되며, 이상값이 통계적으로 추론된 값으로 대체되는 수정된 시계열이 생성된다[43].

- 다층 퍼셉트론 기반 방법 (multi-layer perceptron-based, MLP): 이는 입력 데이터와 출력 데이터 간의 복잡한 매핑을 학습하기 위해 최소 하나의 은닉층을 포함한 여러 층의 노드로 구성된 피드포워드 인공 신경망의 한 유형이다. 이 MLP 기반 방법은 시계열 내의 기본 패턴과 종속성을 학습하도록 MLP를 훈련하는 과정을 포함한다. 훈련이 완료되면 MLP는 누락되거나 비정상적인 값을 예측하고, 통계적으로 추론된 값으로 이러한 공백을 채워 시계열을 효과적으로 보정할 수 있다 [40]].
- 재귀 신경망 기반 방법 (recurrent-neural-network based, RNN): 이 방법은 피드백 연결을 사용하여 순차적 데이터를 처리하도록 설계된 신경망의 한 유형으로, 순서 내 이전 단계의 정보를 기억할 수 있다. 이 RNN 기반 방법은 시간 시리즈에 내재된 시간적 의존성과 패턴을 학습하도록 RNN을 훈련하는 것을 포함한다. 훈련이 완료되면 RNN은 누락되거나 이상적인 값을 복원할 수 있다 [40] .
- 생성적 적대적 네트워크 (Generative adversarial network, GAN): 두 가지 주요 구성 요소인 생성기(generator)와 판별기(discriminator)로 구성된 딥러닝 모델이다. 생성기의 목표는 실제 데이터의 분포와 유사한 새로운 데이터 샘플을 생성하는 것이며, 판별기의 목표는 생성된 데이터와 실제 데이터를 구분하는 것이다. 훈련 과정에서 생성기와 판별기는 적대적 훈련을 통해 가중치를 지속적으로 최적화하며, 결국 생성기가 매우 현실적인 합성 데이터를 생성할 수 있도록 한다. 이 과정은 센서의 측정 값에 발생하는 다양한 이상 현상, 예를 들어 데이터 손실이나 양의 부족 등을 효과적으로 보정할 수 있다 [41].

C.2.3 이상 보정 방법으로 수정 가능한 이상 유형

C.2.2 절에서 설명된 바와 같이, 각 이상 보정 방법은 하나 이상의 유형의 이상 현상을 해결할 수 있다. 각 이상 보정 방법이 수정할 수 있는 이상 현상 유형은 표 C-2에 요약되어 있다.

표 C-2: 이상 유형별 적용 가능한 이상 보정 방법

No	Type of **Repair Method**	**Repair Method**	**Repairable Anomaly Type**
1	Basic	Referenced data comparison	Offset
2		Mean/median/mode imputation	Data loss
3		Interpolation replacement	Trim, Data loss, Different resolution, Inconsistent frequency
4		Resampling	Inconsistent frequency
5		Time synchronisation	Incorrect timestamp
6		Timestamp interpolation	Incorrect timestamp
7	Statistical	Least squares	Drift, Noise, Shift, Drop or rise
8		Polynomial fitting	Shift, Drop or rise
9		Exponential fitting	Drop or rise
10		Dempster-Shafer	Dissimilarity
11	Digital signal processing	Kalman filtering	Offset, Noise
12		Low-pass filtering	Noise

No	Type of **Repair Method**	**Repair Method**	**Repairable Anomaly Type**
13		Band-pass filtering	Noise
14		High-pass filtering	Drift, Noise
15		Median filtering	Spike
16		Weighted moving window filtering	Spike
17		Adaptive filtering	Spike
18		Wavelet transform	Drift, Spike
19		Empirical mode decomposition	Drift, Noise
20		Compressed sensing	Stuck
21		Fourier transform	Bound oscillation
22		Time alignment based on dynamic time w arping	Inconsistent timestamp
23	Machine learning	Neighbour-based	Data loss
24		Constraint-based	Data loss
25		Regression-based	Data loss, Shift, Stuck
26		Matrix factorization-based	Data loss
27		Expectation-maximization-based	Data loss
28		Multi-layer perceptron-based	Offset, Trim, Data loss, Dissimilarity
29		Recurrent-neural-network-based	Noise, Drop or rise
30		Generative adversarial network	Lack of amount, Data loss

D. 센서 데이터 정제 사례

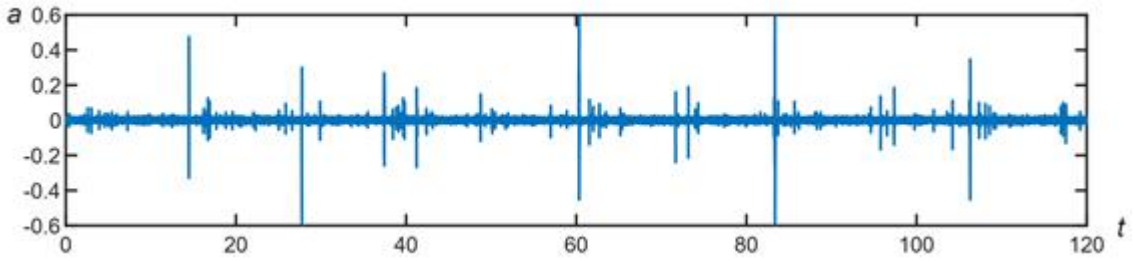
D.1 사례 개요

CRCHI (China Railway Construction Heavy Industry Group Co., Ltd) 는 지능형 터널 건설 장비와 농업 장비의 연구, 설계, 제조 및 서비스를 통합한 대형 전문 기업이다. 이 회사의 대표적인 제품은 터널 보링 머신(TBM)으로, 운영 상태를 모니터링하기 위해 800개 이상의 센서가 장착되어 있다. 이들 센서 중에서 TBM에 설치된 진동 가속도 센서, 변형(스트레인) 센서, 실린더 스트로크 센서는 장비 모니터링을 위한 주요 센서이며, 그 데이터는 데이터 분석 및 활용에서 자주 사용된다. 따라서 본 절에서는 이 세 가지 센서로부터 수집된 데이터의 정제 사례를 제시한다. 이 가운데, D.2와 D.3은 개별 센서에 대한 데이터 이상 정제 사례이며, D.4는 다중 센서에 대한 사례이다.

D.2 TBM 진동 가속도 센서 데이터 정제

D.2.1 진동 가속도 센서 데이터 정제 수행

TBM의 가혹한 작업 환경으로 인해 진동 가속도 센서 (vibration acceleration sensor) 데이터의 품질은 데이터 분석이나 활용에 직접 사용할 만큼 충분히 높지 않다. 그림 D-1은 120초의 시간 구간 동안 수집된 진동 가속도 센서 데이터의 예시를 보여주는데, 여기에서 스파이크 이상값이 자주 발견된다. 진동 가속도 데이터의 품질을 향상시키기 위해서는 스파이크 이상값을 보정하여 정제하는 과정이 필요하다.



Key
a: TBM의 진동 가속도, 중력 가속도로 표현됨
t: 데이터 수집 시간, 초 단위로 표현됨

그림 D-1: 원시 진동 가속도 센서 데이터(스파이크 포함)

진동 가속도 센서 데이터 정제 수행(Perform Vibration Acceleration Sensor Data Cleansing) 프로세스(A0)의 컨텍스트 다이어그램은 그림 그림 D-2에 제시되어 있다. ISO 8000-210에서 언급된 바와 같이, 스파이크 이상값은 품질 특성 중 정확성에 영향을 미친다. 따라서 이 예시에서는 스파이크 이상을 보정함으로써 정확성를 개선할 수 있다. ISO 8000-220에 따르면, 데이터 세트의 정확성에 대한 데이터 품질 측정값은 X = A/B로 정의되며, 여기서 A는 정확한 것으로 평가된 데이터 값의 개수이고, B는 평가된 전체 데이터 값의 개수이다. 그림 D-2의 원시 센서 데이터 예시에서 정확성는 99.72%이다. 유효한 과거 센서 데이터를 기반으로 데이터 품질 요구사항은 X ≥ 99.8%로 설정되며, 이는 센서 데이터의 정확성가 99.8% 이상이어야 함을 의미한다.

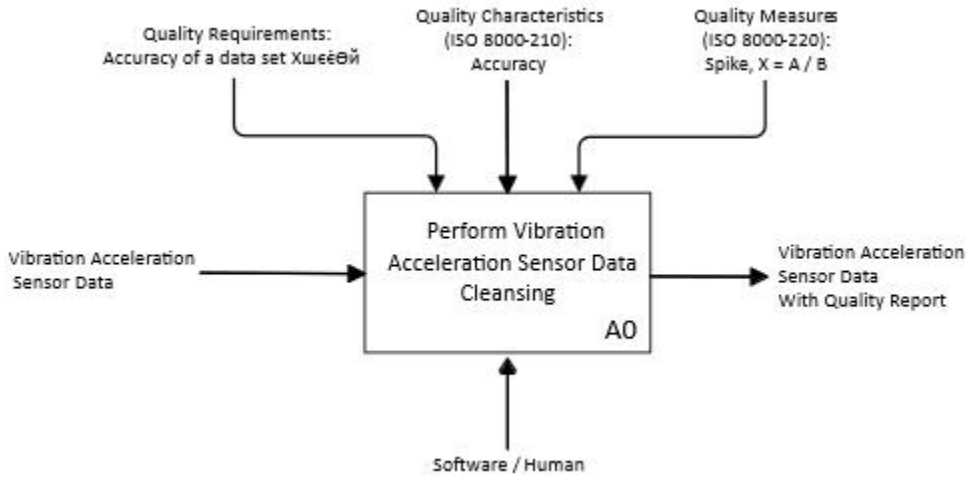


그림 D-2: 진동 가속도 센서 데이터 정제 수행을 위한 A-0 컨텍스트 다이어그램(모델 다이어그램 A0)

진동 가속도 센서 데이터 정제 수행은 그림 D-3 에 제시되어 있으며, 이는 다음의 세 가지 활동으로 구성된다:

1) A1
   - 과거 진동 가속도 센서 데이터의 특성에 따라 데이터 세트 정확성의 품질 요구사항을 X ≥ 99.8%로 결정
   - TBM에서 수집된 진동 가속도 데이터에 대한 측정 계획(Prepare Vibration Acceleration Data Measurement Plan)을 수립

2) A2
   - 측정 계획에 근거하여 이상 탐지 모델을 도출하고 진동 가속도 데이터의 품질을 측정 (Measure Vibration Acceleration Data Quality)
   - 만약 센서 데이터의 품질을 개선할 기회가 없다면, 정제를 중단하고 품질 보고서가 포함된 센서 데이터를 출력
   - 그렇지 않은 경우, 품질 향상을 위해 활동 A3로 진행

3) A3
   - 데이터 보정 계획을 수립하고 적절한 보정 방법을 사용하여 데이터를 보정
   - 정제된 센서 데이터를 출력 (Improve Vibration Acceleration Data Quality).

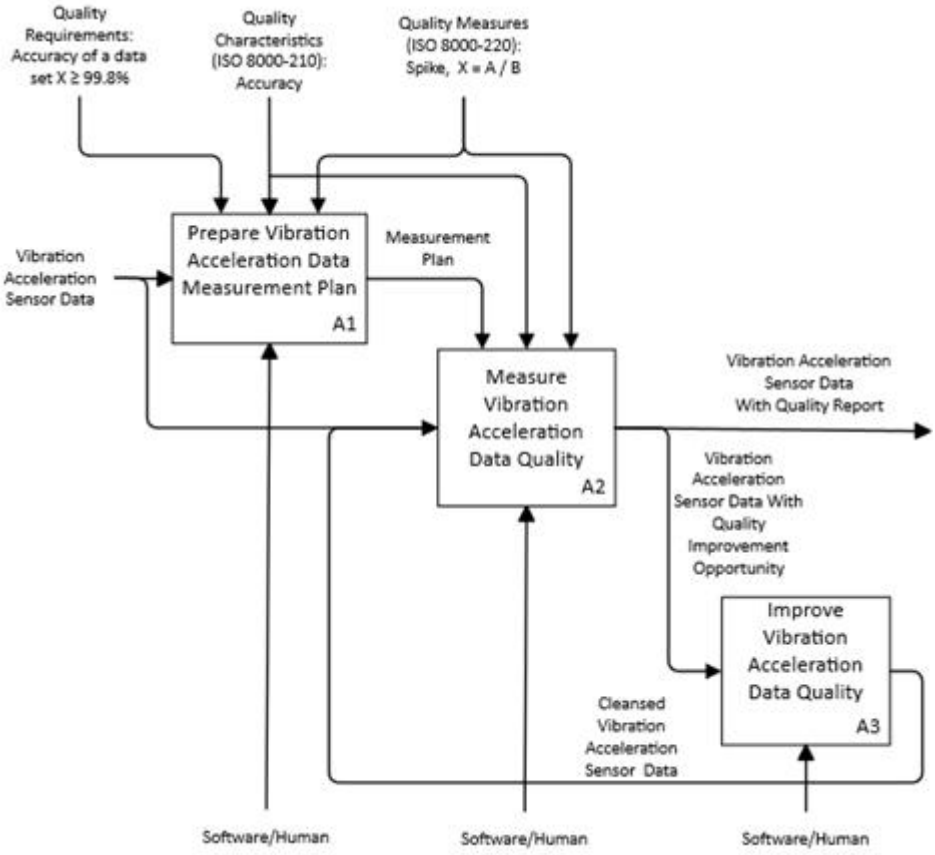


그림 D-3: 진동 가속도 센서 데이터 정제 수행 (모델 다이어그램 A0)

D.2.2 진동 가속도 데이터 측정 계획 수립 (A1)

그림 D-4에서 보듯이, 진동 가속도 데이터 측정 계획 수립((Prepare Vibration Acceleration Data Measurement Plan, A1)에 의해 데이터 품질 측정 계획이 제공되며, 이는 다음의 세 가지 하위 활동으로 구성된다:

1) A11
   - TBM에서 수집된 진동 가속도 센서 데이터에 대해, 데이터 세트 정확성 품질 요구사항 X ≥ 99.8%로 표현되는 진동 가속도 데이터 품질 목표를 수립 (Establish Vibration Acceleration Data Quality Goal)

2) A12
   - 품질 목표와 TBM에서 수집된 과거 진동 가속도 센서 데이터의 특성에 따라 데이터 프로파일링을 수행하여 TBM 진동 가속도 센서 데이터의 특성을 도출(Perform Vibration Acceleration Data Profiling)
   - 통계 분석 결과, 진동 가속도 센서 데이터는 정규분포 규칙을 준수.

3) A13
   - 과거 센서 데이터에서 얻어진 정규분포 규칙을 참조하여 효과적인 데이터 품질 측정 계획을 수립 (Develop Vibration Acceleration Data Measurement Plan)

- 데이터 세트 정확성의 품질 목표 X ≥ 99.8%와 정규분포 패턴에 근거하여, 3시그마를 벗어난 데이터는 스파이크로 정의

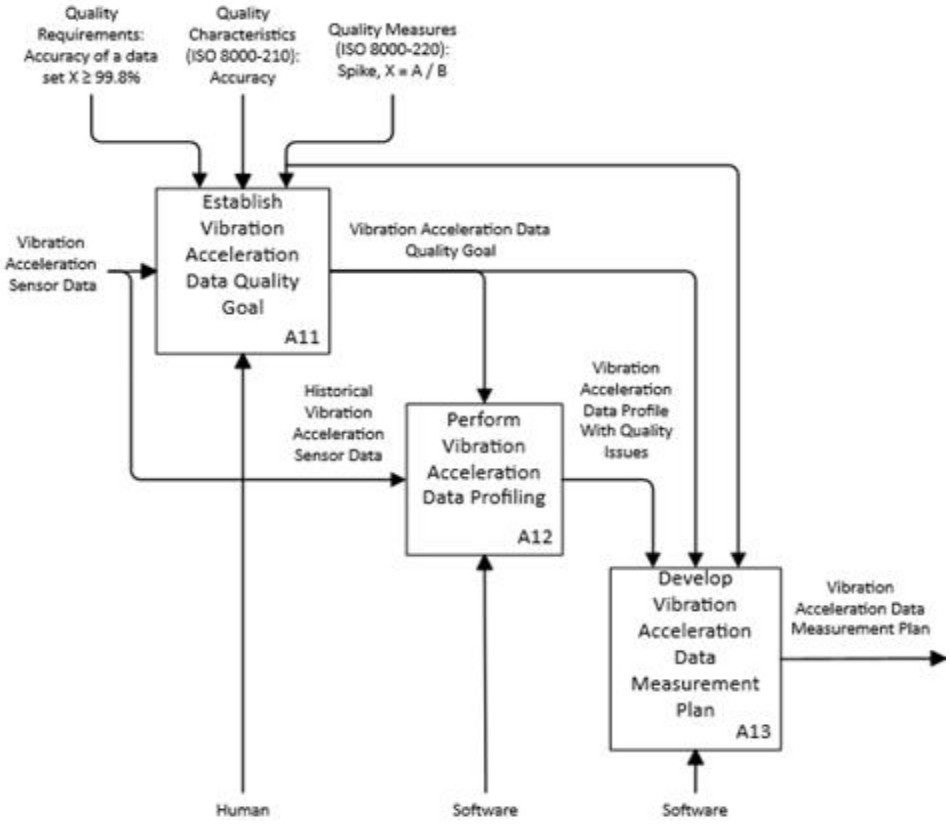


그림 D-4: 진동 가속도 데이터 측정 계획 수립 (모델 다이어그램 A1)

D.2.3 진동 가속도 데이터 품질 측정 (A2)

그림 D-5에서 보듯이, 진동 가속도 센서 데이터의 품질은 진동 가속도 데이터 품질 측정(Measure Vibration Acceleration Data Quality, A2)에 의해 평가되며, 이는 다음 두 가지 하위 활동으로 구성된다:

1) A21

- 측정 계획에 따라 스파이크 이상값을 탐지하기 위한 이상 탐지 모델을 도출(Derive Anomaly Detection Model)
- 정규분포 변수가 허용 범위 내에 있는지를 판단하기 위해 임계값 μ±3σ를 설정
  i) μ는 2초 슬라이딩 시간 창(40,000 데이터 포인트) 내 데이터 시퀀스의 평균값
  ii) σ는 동일한 시간 창 내 데이터 시퀀스의 표준편차

2) A22

- 데이터 세트의 품질 측정 정확성가 품질 요구사항을 충족한다면, 추가적인 데이터 정제는 필요하지 않다. 그렇지 않다면, 진동 가속도 데이터에서 정제가 필요한 스파이크 이상값을 이상 탐지 모델로 탐지한다. 데이터 세트의 정확성가 X = 99.72%이므로, 센서 데이터는 품질 요구사항 X ≥ 99.8%를 충족하지 않는다. 따라서 원시 센서 데이터에서 스파이크 이상값을 탐지하고, 스파이크로 인한 부정확성를 계산한다. 스파이크로 인한 부정확성는 X_S = 0.28%이며, 스파이크 이상값은 개선 가능하다. 따라서 품질 개선 기회가 있는 센서 데이터는 다음 활동인 진동 가속도

데이터 품질 개선(A3)으로 진행된다(Find Vibration Acceleration Data Quality Improvement Opportunity).

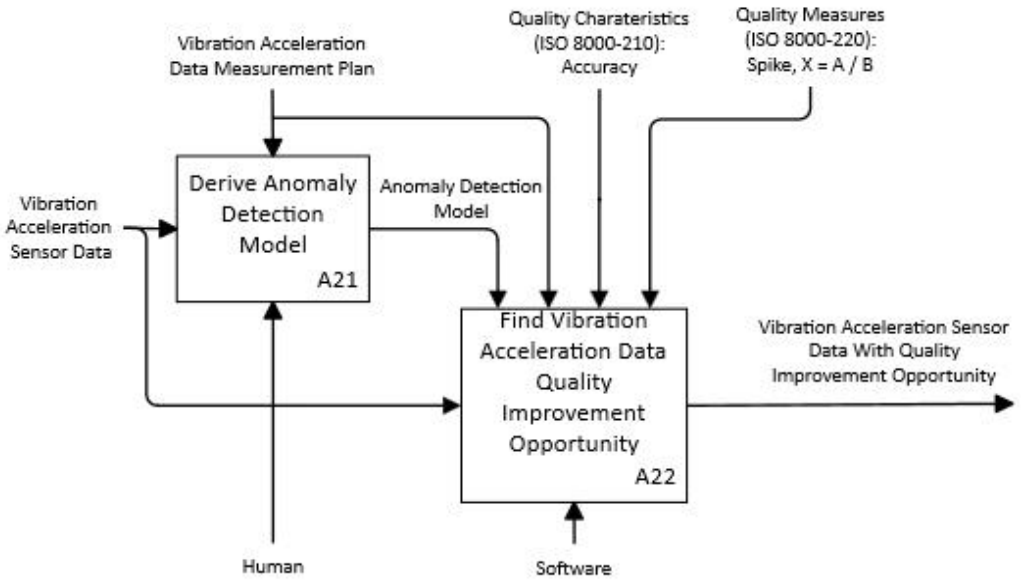


그림 D-5: 진동 가속도 데이터 품질 측정 (모델 다이어그램 A2)

D.2.4 진동 가속도 데이터 품질 개선 (A3)

그림 D-6에서 보듯이, 진동 가속도 데이터의 품질은 진동 가속도 데이터 품질 개선((Improve Vibration Acceleration Data Quality, A3)에 의해 향상되며, 이는 다음 세 가지 하위 활동으로 구성된다:

1) A31
   - 중앙값 필터링(median filtering) 방법에 기반하여 진동 가속도 센서 데이터에 대한 데이터 보정 계획을 수립한다(Establish Vibration Acceleration Data Repair Plan). 이 방법에서는 스파이크 이상값을, 스파이크가 발생한 시점 전후 1초(총 2초 슬라이딩 시간 창) 구간 내 평균값으로 대체한다.

2) A32
   - 진동 가속도 센서 데이터 책임자로부터 보정 계획에 대한 승인을 받는다(Confirm Vibration Acceleration Data Repair Plan).

3) A33
   - 데이터 보정 계획이 실행 가능하다고 확인되면, 데이터 보정을 수행하고 정제된 진동 가속도 센서 데이터를 생성한다(Execute Vibration Acceleration Data Repair).

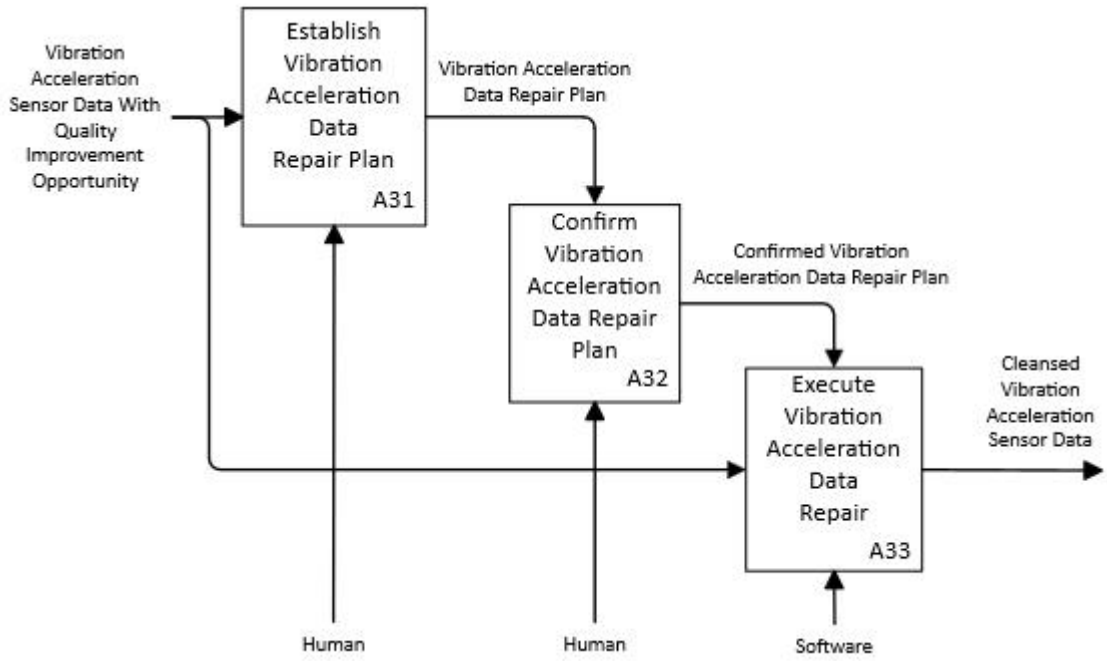
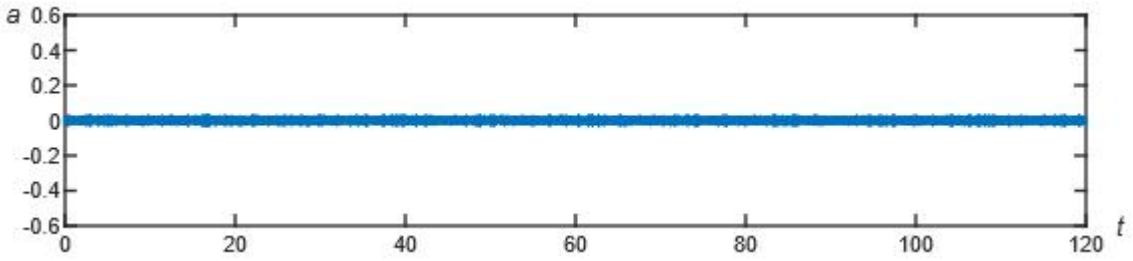


그림 D-6: 진동 가속도 데이터 품질 개선 (모델 다이어그램 A3)

모든 스파이크 이상값을 보정한 후, 품질 개선 기회 확인(A22)을 다시 수행하여 정제된 진동 가속도 데이터가 품질 요구사항을 충족하는지 확인한다. 데이터 세트의 정확성 품질 측정값은 99.91%로 향상되었으며, 정제된 진동 가속도 데이터는 품질 요구사항 X ≥ 99.8%를 충족한다. 따라서 품질 결과 보고(A23)가 수행되어, 품질 보고서를 포함한 품질 개선된 진동 가속도 데이터가 최종 결과로 생성된다.

정제 이후 획득한 품질 개선 진동 가속도 데이터는 그림 D-7에 제시되어 있다.



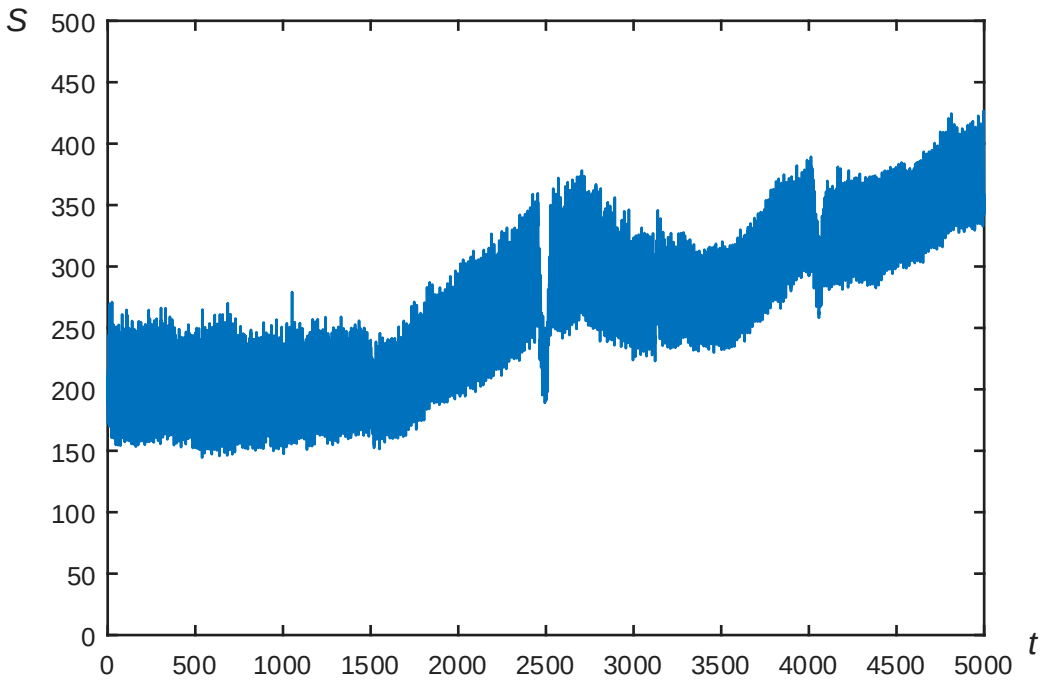
Key
a: TBM의 진동 가속도, 중력 가속도로 표현됨
t: 데이터 수집 시간, 초 단위로 표현됨

그림 D-7: 데이터 정제 이후의 진동 가속도 센서 데이터

D.3 TBM 변형률 센서 데이터 정제

D.3.1 TBM 변형률 센서 데이터 정제 수행

터널 굴착 작업 동안, TBM은 가혹한 작업 환경에서 사용되며 종종 큰 하중을 받게 된다. 이러한 이유로 TBM의 구조적 변형률(Strain)은 변형률 센서를 통해 모니터링된다. 온도 변화와 같은 요인으로 인해 변형률 신호가 드리프트(drift) 되는 것을 경험할 수 있으며, 이로 인해 데이터 품질이 저하된다. 안정적인 굴착 조건에서 TBM의 오프라인 변형률 센서 데이터가 지속적으로 증가하는 시퀀스를 보이며, 명백한 드리프트 이상이 발생한 사례가 그림 D-8에 제시되어 있다. 한편, 2500초와 4200초에서 나타나는 스파이크는 이상이 아니라, 건설 과정 중 나타나는 정상적인 동적 변형률 응답이다.



Key
s: TBM의 변형률(마이크로 변형률 단위)
t: 데이터 수집 시간, 초 단위로 표현됨

그림 D-8: 원시 변형률 센서 데이터(드리프트 포함)

변형률 센서 데이터 정제 수행(Perform Strain Sensor Data Cleansing) 프로세스의 컨텍스트 다이어그램은 그림 D-9에 표시된다. ISO 8000-210에서 언급된 바와 같이, 드리프트 이상은 품질 특성인 정확성에 영향을 미친다. 따라서 이 예제에서는 드리프트 이상을 수정하여 정확성를 향상시킬 수 있다. ISO 8000-220에 따르면, 데이터 세트의 정확성에 대한 데이터 품질 측정은 X = A/B이며, 여기서 A는 정확한 것으로 평가된 데이터 값의 개수이고, B는 평가된 모든 데이터 값의 개수이다. 예제의 원시 변형률 센서 데이터의 정확성는 56.68%이다. 유효한 과거 변형률 센서 데이터를 기반으로 데이터 품질 요구사항은 X ≥ 99%로 설정되며, 이는 센서 데이터의 정확성가 99%

이상이어야 함을 의미한다. ISO 8000-210에 따르면, 드리프트 이상은 데이터 품질 특성 중 정확성에 영향을 미친다. 따라서 본 예제에서는 드리프트 이상을 보정하여 정확성을 향상시킨다. ISO 8000-220에 의하면, 데이터 세트의 정확성은 A/B로 정의되며, A는 정확한 데이터 값의 수, B는 전체 평가된 데이터 값의 수이다. 그림 D-8의 원시 변형률 센서 데이터 정확성는 56.68%이다. 유효한 과거 변형률 센서 데이터에 근거하여 데이터 품질 요구사항은 $X \geq 99\%$으로 설정되며, 이는 센서 데이터 정확성가 99% 이상이어야 함을 의미한다.

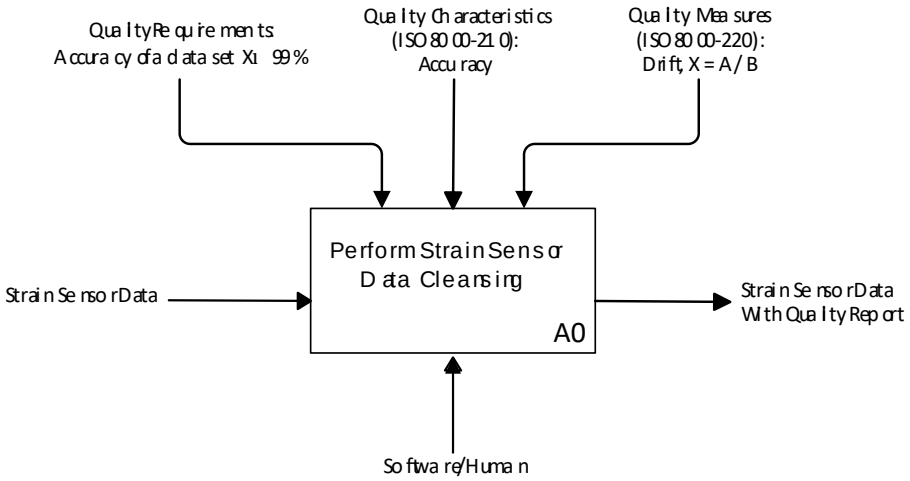


그림 D-9: 변형률 센서 데이터 정제 수행을 위한 A-0 컨텍스트 다이어그램

변형률 센서 데이터 정제 수행(Perform Strain Sensor Data Cleansing) 프로세스는 그림 D-10에 표시되어 있으며 다음과 같은 3가지 활동으로 구성된다:

1) A1
   - TBM에서 수집된 변형률 센서 데이터 세그먼트에 대해 품질 측정 계획 수립(Prepare Strain Data Measurement Plan)

2) A2
   - 측정 계획에 따라 이상 탐지 모델을 도출하고 품질을 측정(Measure Strain Data Quality)
   - 개선 여지가 없으면 종료하고 보고서를 생성
   - 개선 가능성이 있으면 A3로 진행

3) A3
   - 수립된 보정 계획에 따라 데이터 정제를 수행하고 정제된 센서 데이터를 산출(Improve Strain Data Quality)

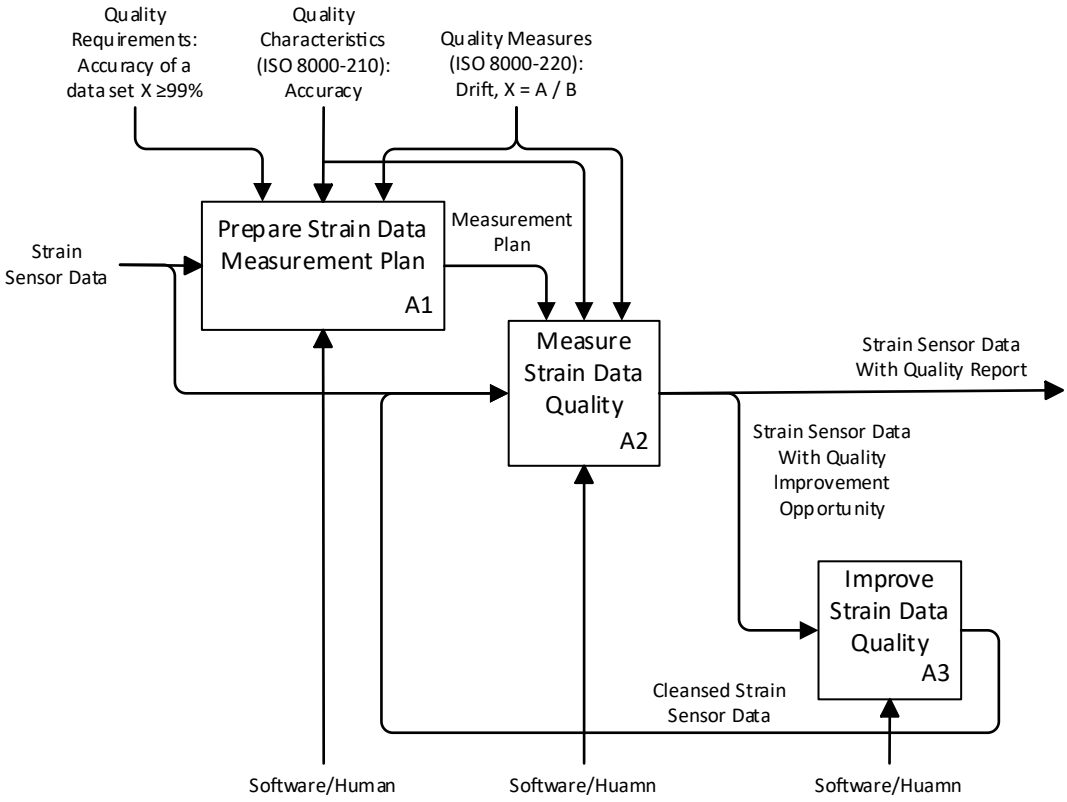


그림 D-10: 변형률 센서 데이터 정제 수행(모델 다이어그램 A0)

D.3.2 변형률 데이터 측정 계획 수립 (A1)

그림 D-11에서 보이는 바와 같이, 데이터 정제 전에 다음과 같은 활동을 통해 데이터 품질 측정 계획 수립(Prepare Strain Data Measurement Plan)을 한다:

1) A11
   - TBM에서 수집한 변형률 센서 데이터를 대상으로 데이터 품질 목표를 설정한다(Establish Strain Data Quality Goal). 즉, 연속된 2초 슬라이딩 시간 창 세 구간에서의 변형률 센서 데이터 평균값의 합이 기준값의 30με (3*10 마이크로 변형)을 초과하지 않아야 하며, 변형률 센서 데이터의 품질 목표는 정확성 X ≥ 99%로 한다.

2) A12
   - 품질 목표와 TBM에서 수집한 과거 변형률 센서 데이터의 특성을 고려하여 데이터 프로파일링을 수행하고, 변형률 센서 데이터의 특성을 파악(Perform Strain Data Profiling).
     i) 온도 변화로 인해 발생하는 드리프트 이상값은 변형률 센서 신호에 서서히 나타나며 지속적인 상승 또는 하강으로 표현된다.
     ii) 예를 들어, 드리프트 이상값은 1834초 시점부터 변형률 센서 데이터의 연속적 증가로 나타날 수 있다. 드리프트는 다음 조건으로 정의된다: $|C(t)_i - F(t)_i| > 10\mu\varepsilon$, $i=1,2,3,\ldots$ 여기서 $C(t)_i$는 현재 2초 슬라이딩 시간 창에서 데이터 시퀀스의 평균값이고, $F(t)_i$ 는 현재 시간 창 20분 전 첫 번째 2초 시간 창에서 데이터 시퀀스의 평균값이다. 만약 $C(t)_i$, $C(t)_{(i+1)}$, $C(t)_{(i+2)}$ 가 모두 $|C(t)_i - F(t)_i| > 10\mu\varepsilon$ 조건을 만족한다면 드리프트가 발생한 것으로 간주한다. 이는 30

마이크로 변형(해당 응력은 약 6MPa에 해당하며 구조적 응력 측정 오차 허용 한계를 초과함)이므로 허용 불가하다.

3) A13

- 과거 센서 데이터에서 얻어진 드리프트 이상값 판정 방법을 참조하여 품질 측정 계획을 수립(Develop Strain Data Measurement Plan).
- A12에서 정의된 기준을 만족하는 센서 데이터는 드리프트로 간주하며, 드리프트로 인한 정확성는 X ≥ 99%로 한다.

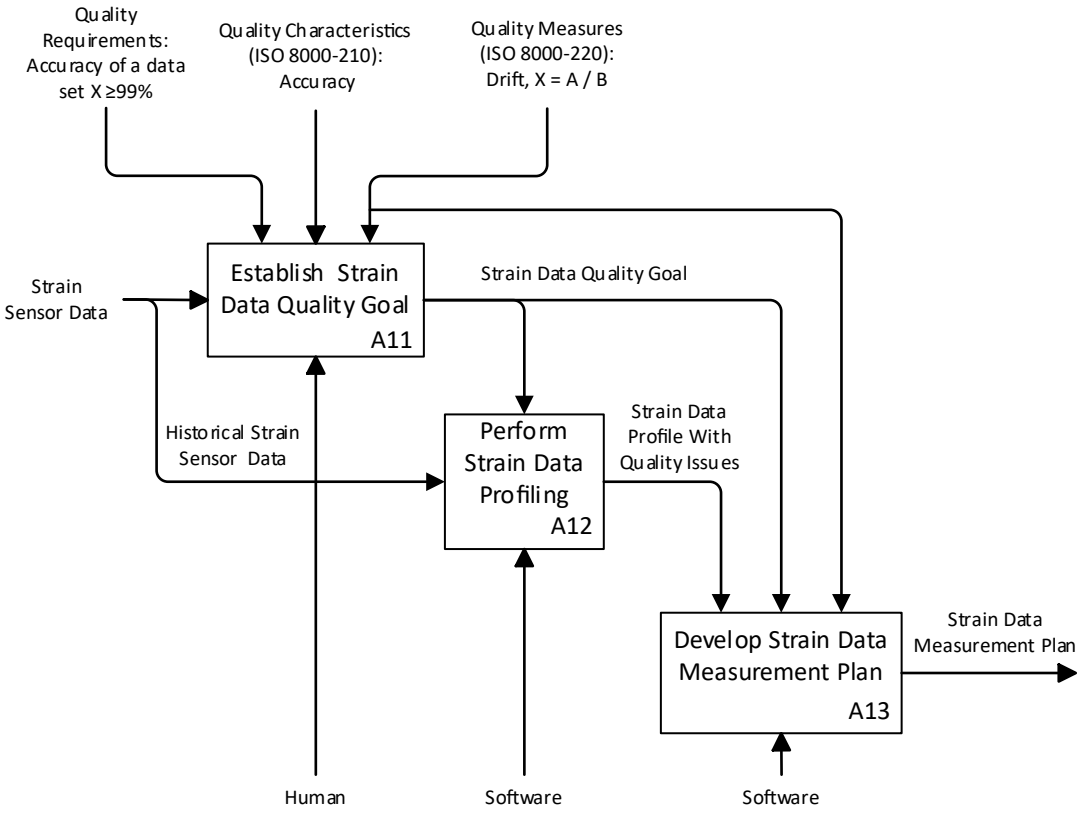


그림 D-11: 변형률 데이터 측정 계획 수립 (모델 다이어그램 A1)

D.3.3 변형률 데이터 품질 측정 (A2)

그림 D-12와 같이, 변형률 센서 데이터의 품질은 측정은 변형률 데이터 품질 측정((Measure Strain Data Quality, A2)에 의해 측정되며, 이것은 다음 두 개의 하위 활동으로 구성된다:

1) A21

- A12에서 정의된 드리프트 판단 방법에 기반하여 이상 탐지 모델을 도출하고 품질 측정값 X를 계산(Derive Anomaly Detection Model).

2) A22

- 데이터 세트의 품질 측정 정확성가 품질 요구사항을 충족한다면, 추가적인 데이터 정제는 필요하지 않다. 그렇지 않은 경우, 이상 탐지 모델을 사용하여 정제가 필요한 변형률 데이터의 드리프트 이상값을 탐지한다.

- 데이터 세트의 정확성가 X = 36.68%이므로, 센서 데이터는 품질 요구사항인 X ≥ 99%를 충족하지 못한다. 따라서 원시 센서 데이터에서 드리프트 이상값을 탐지하고, 드리프트로 인한 품질 측정 부정확성를 계산한다. 드리프트로 인한 부정확성는 Xd=63.32이며, 드리프트 이상값은 개선 가능하다. 따라서 품질 개선 기회가 있는 센서 데이터는 다음 활동인 변형률 데이터 품질 개선 (A3)으로 진행된다(Find Strain Sensor Data Quality Improvement Opportunity).

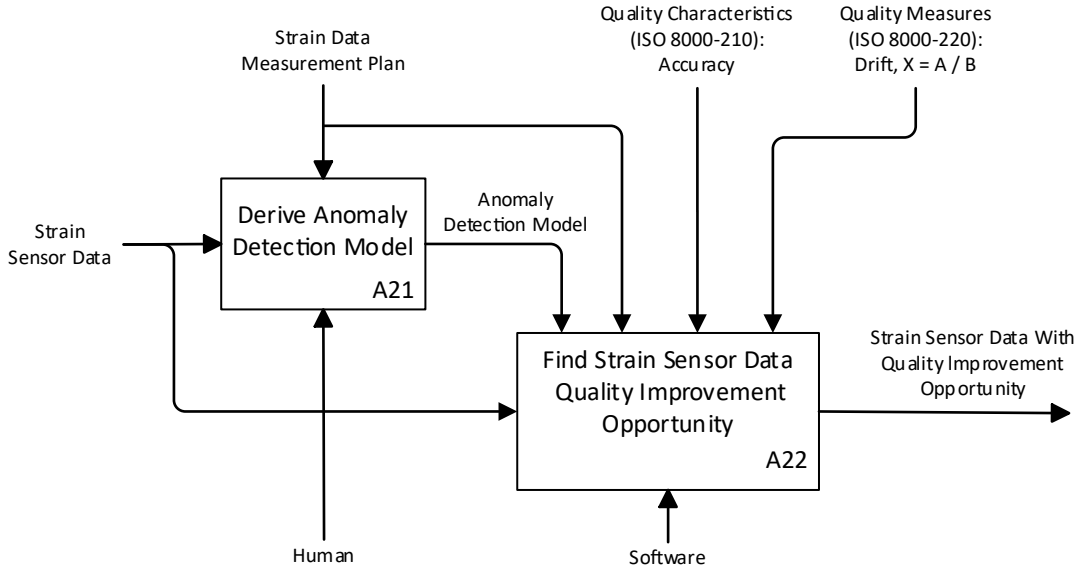


그림 D-12: 변형률 데이터 품질 측정(모델 다이어그램 A2)

D.3.4 변형률 데이터 품질 개선 (A3)

그림 D-13에서 보이는 바와 같이, 변형률 데이터의 품질은 변형률 데이터 품질 개선 (Improve Strain Data Quality, A3) 활동을 통해 향상되며, 이는 다음의 세 가지 하위 활동으로 구성된다:

1) A31
   - 변형률 센서 데이터를 위한 데이터 보정 계획을 수립(Establish Strain Data Repair Plan).
   - 드리프트 이상값을 제거하기 위해 경험적 모드 분해(EMD, Empirical Mode Decomposition) 방법을 사용.
   - 우선, 품질 개선 기회가 있는 데이터를 EMD에 적용하여 고유 모드 함수(IMFs, Intrinsic Mode Functions)를 도출한다. 이후 드리프트 경향은 일반적으로 저주파 신호이므로, IMFs에서 저주파 성분을 제거한다. 마지막으로 남은 IMFs를 재구성하여 데이터 정제를 완료한다.

2) A32
   - 변형률 센서 데이터의 책임자로부터 보정 계획에 대한 동의를 얻는다(Confirm Strain Data Repair Plan).

3) A33
   - 확정된 데이터 보정 계획에 따라 데이터 보정을 실행(Execute Strain Data Repair)하고, 정제된 변형률 센서 데이터를 생성한다.

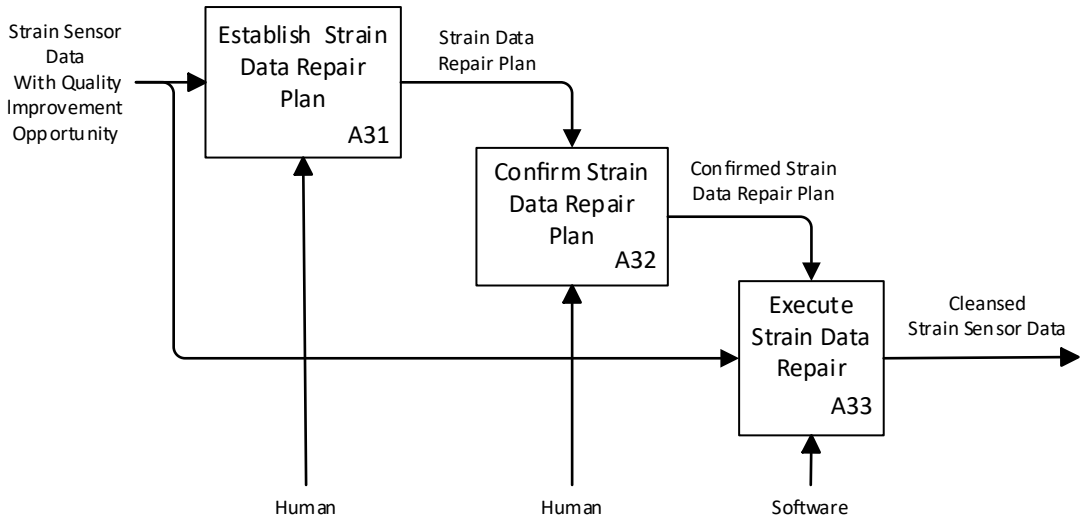
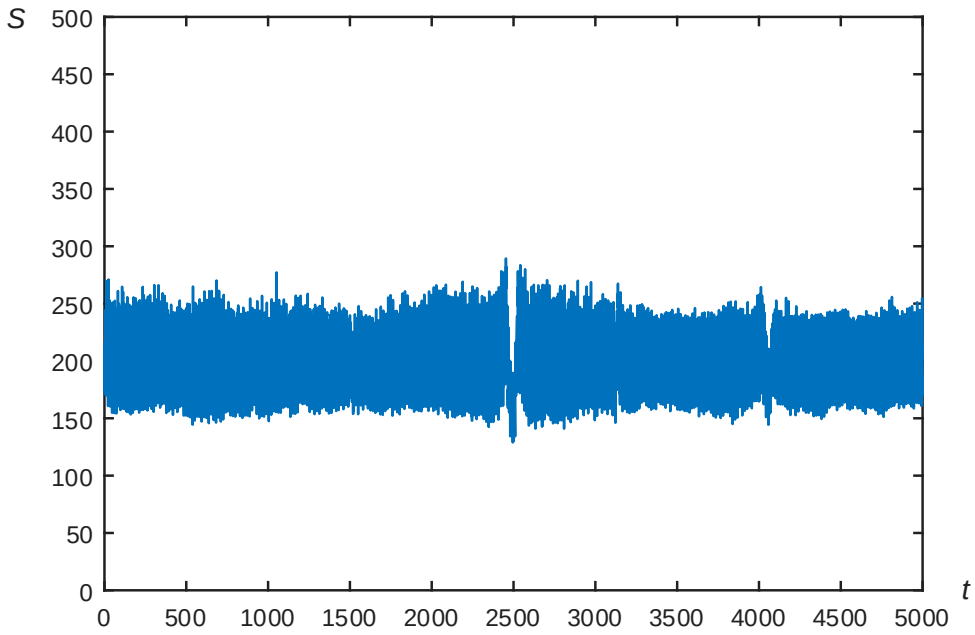


그림 D-13: 변형률 데이터 품질 개선(모델 다이어그램 A3)

드리프트 이상값을 보정한 후, 품질 개선 기회 탐색 (A22)를 다시 수행하여 정제된 변형률 데이터가 품질 요구사항을 충족하는지 확인한다. 데이터 세트의 정확성 품질 측정값이 99.92%로 향상되었으므로, 정제된 변형률 데이터는 품질 요구사항 X ≥ 99%를 만족한다. 따라서 품질 결과 보고 (A23)를 수행하여, 품질 보고서를 포함한 품질 개선 변형률 데이터가 결과물로 생성된다.

정제 후 얻어진 품질 개선 변형률 데이터는 그림 D-14에 나타나 있다.



Key

s: TBM의 변형률(마이크로 변형률 단위)

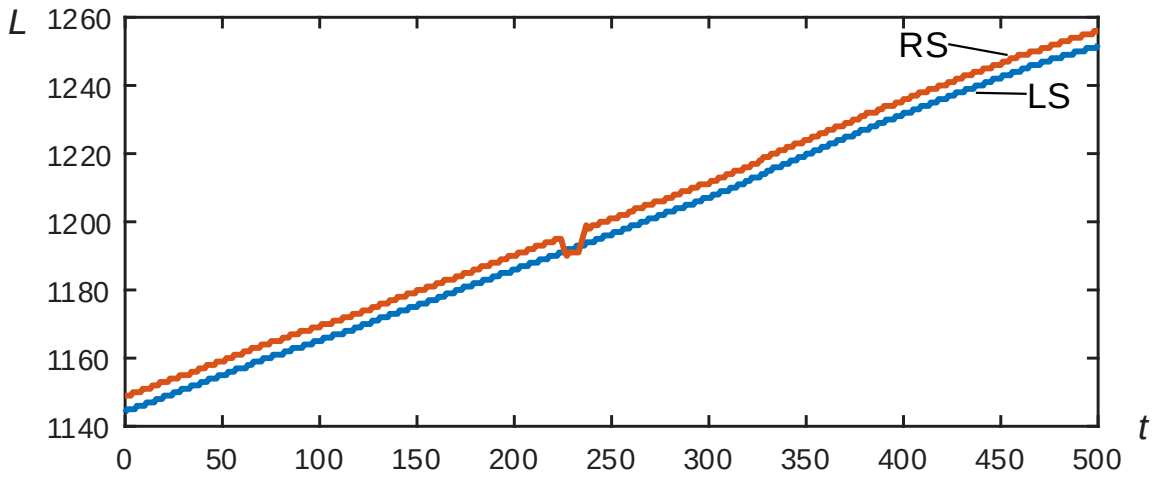
t: 데이터 수집 시간, 초 단위로 표현됨

그림 D-14: 데이터 정제 후의 변형률 센서 데이터

D.4 TBM 추진 실린더 스트로크 센서의 다중 센서 데이터 정제

D.4.1 스트로크 센서 데이터 정제 수행

터널 굴착 작업 중, TBM은 종종 굴진 방향을 조정할 필요가 있다. 이 경우 TBM은 굴진 방향의 각도를 조정함으로써 조향을 정확히 수행하기 위해 모든 추진 실린더의 스트로크를 모니터링하고 조정한다. 이러한 실린더는 쌍(pair)으로 구성되며, 각 쌍은 하나의 스트로크 센서를 공유한다. 그러나 이러한 스트로크 센서는 때때로 외부 간섭의 영향을 받아 측정 데이터에 편차가 발생하고 데이터 품질이 저하될 수 있다. 그림 D.15는 TBM이 좌측으로 방향을 전환할 때 좌측 및 우측 실린더의 스트로크 센서 데이터를 보여준다. 두 센서 데이터 세트 간의 규칙에 따르면, 우측 실린더의 스트로크가 좌측 실린더의 스트로크보다 커야 한다. 그러나 이 예시에서는 224초에서 237초 구간 동안 우측 실린더(RS)의 스트로크가 좌측 실린더(LS)의 스트로크보다 작게 나타나는 데이터 이상값이 발생하였다. 따라서 데이터 품질을 향상시키기 위해서는 우측 실린더의 스트로크 센서 데이터를 정제해야 한다.



Key

L	추진 실린더 스트로크, 밀리미터(mm) 단위로 표시
t	데이터 수집 시간, 초(second) 단위로 표시
LS	좌측 실린더 스트로크
RS	우측 실린더 스트로크

그림 D-15: 원시 스트로크 센서 데이터(규칙 위반 포함)

전체 스트로크 센서 데이터 정제 수행(Perform Stroke Sensor Data Cleansing) 프로세스(A0)의 컨텍스트 다이어그램은 그림 D-16에 나타나 있다. ISO 8000-220에 따르면, 데이터 세트 간 일관성(consistency)은 다음과 같이 정의된다: X=A/B. 여기서 A는 적용 가능한 규칙을 만족하는 데이터 값의 개수이고, B는 평가된 모든 데이터 값의 개수이다. 그림 D-15의 예에서 좌·우 스트로크 센서 데이터 세트 간 일관성은 97.2%이다. 유효한 과거 센서 데이터를 기준으로 데이터 품질 요구사항은 X = 100%로 설정된다.

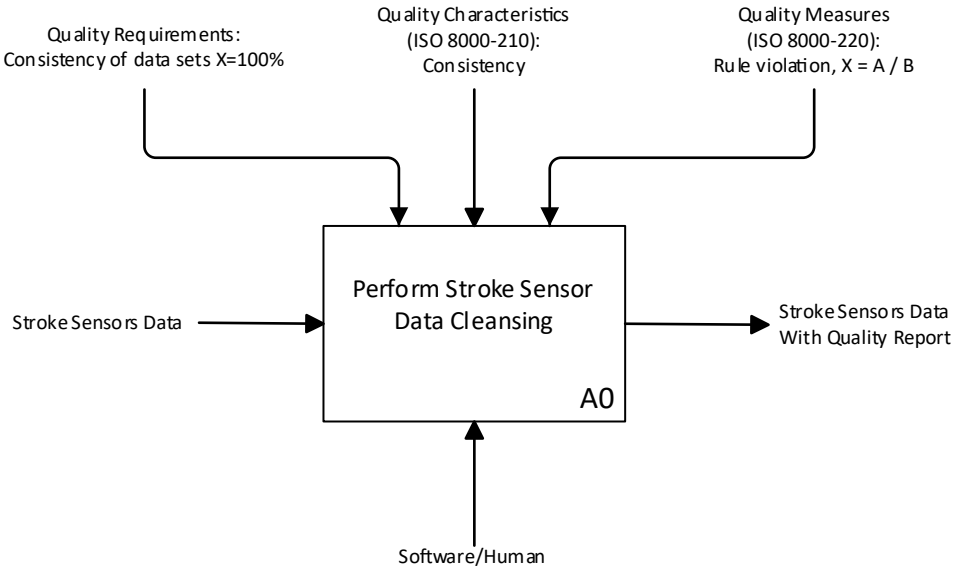


그림 D-16: 스트로크 센서 데이터 정제 수행을 위한 A-0 컨텍스트 다이어그램 (모델 다이어그램 A0)

스트로크 센서 데이터 정제 수행 (Perform Stroke Sensor Data Cleansing) 프로세스는 그림 D-17에 나타나 있으며, 다음 세 가지 활동으로 구성된다:

1) A1
   - 과거 스트로크 센서 데이터의 특성에 따라 데이터 세트 간 일관성(consistency)에 대한 품질 요구사항을 X = 100%로 결정하고, TBM에서 수집한 추진 실린더 스트로크 데이터에 대한 측정 계획을 수립(Prepare Stroke Data Measurement Plan).

2) A2
   - 측정 계획을 기반으로 이상 탐지 모델을 도출하고 스트로크 데이터의 품질을 측정(Measure Stroke Data Quality).
   - 만약 데이터 품질을 개선할 기회가 없다면 정제를 중단하고 품질 보고서가 포함된 센서 데이터를 출력한다.
   - 그렇지 않으면 품질 개선을 위해 활동 A3으로 진행한다.

3) A3
   - 데이터 보정 계획을 수립하고, 적절한 보정 방법을 사용하여 데이터를 보정한 후 정제된 센서 데이터를 출력(Improve Stroke Data Quality).

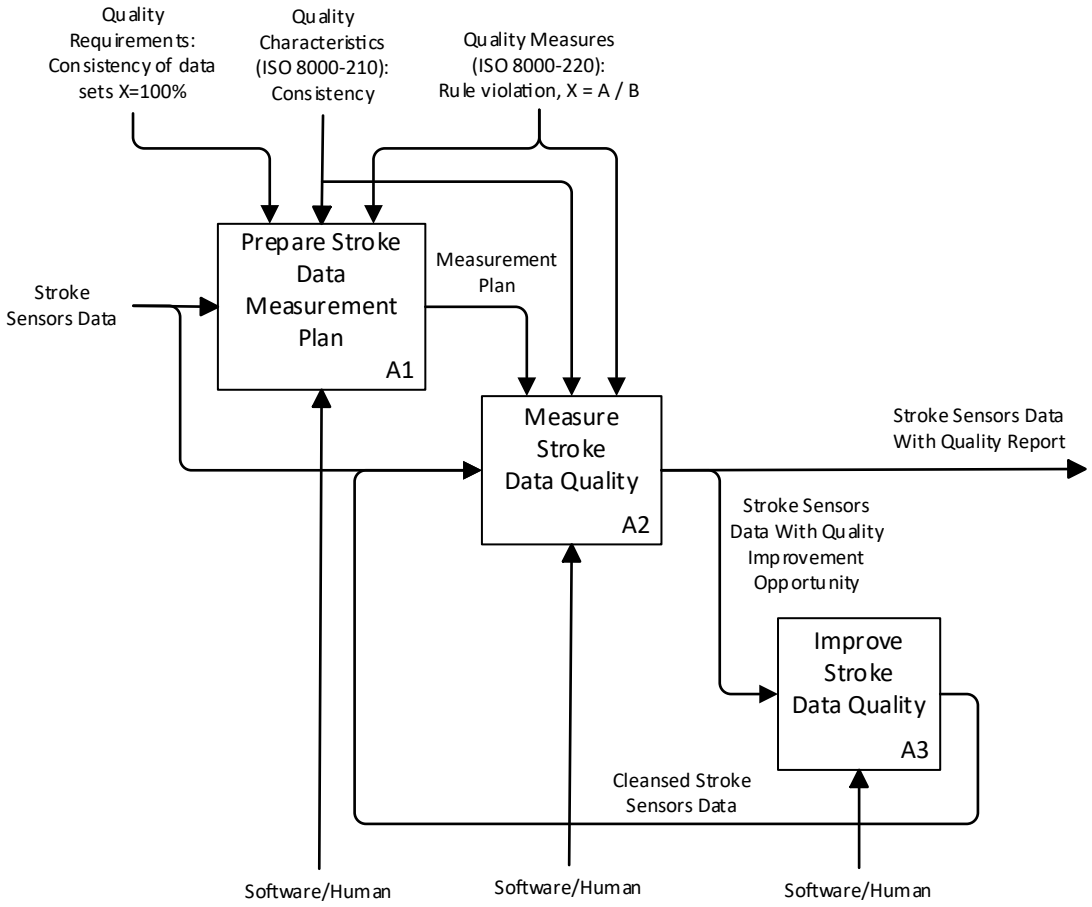


그림 D-17: 스트로크 센서 데이터 정제 수행 (모델 다이어그램 A0)

D.4.2 스트로크 데이터 측정 계획 수립 (A1)

그림 D-18에서 보이는 바와 같이, 스트로크 데이터 측정 계획 수립 (Prepare Stroke Data Measurement Plan, A1) 은 다음 세 가지 하위 활동으로 구성되어 데이터 품질 측정 계획을 제공한다:

1) A11
   - TBM 추진 실린더의 스트로크 센서 데이터, 특히 굴진 방향 전환 시의 데이터를 중심으로 데이터 품질 목표를 설정(Establish Stroke Data Quality Goal).
   - 이 목표는 데이터 세트 간 일관성(consistency)에 대한 품질 요구사항으로, X = 100%로 표현된다. 이는 TBM이 좌측으로 회전할 때 우측 실린더의 스트로크가 좌측 실린더의 스트로크보다 커야 한다는 규칙을 의미

2) A12
   - 품질 목표와 수집된 과거 스트로크 센서 데이터의 특성에 따라 데이터 프로파일링을 수행하여 스트로크 센서 데이터의 특성을 도출(Perform Stroke Data Profiling)
   - TBM이 좌측으로 회전할 때, 우측 실린더의 스트로크는 좌측 실린더의 스트로크보다 커야 하며, 두 값의 차이는 일정한 값으로 유지

3) A13

- 과거 스트로크 센서 데이터로부터 얻어진 조향 실린더의 스트로크 판정 방법을 참조하여 효과적인 데이터 품질 측정 계획을 수립(Develop Stroke Data Measurement Plan).
- 데이터 세트 간 일관성의 품질 목표 X = 100%에 기반하여, 조향 실린더의 스트로크 판정 방법에 따라 좌측과 우측 실린더의 스트로크 센서 데이터는 일정한 범위의 차이를 유지

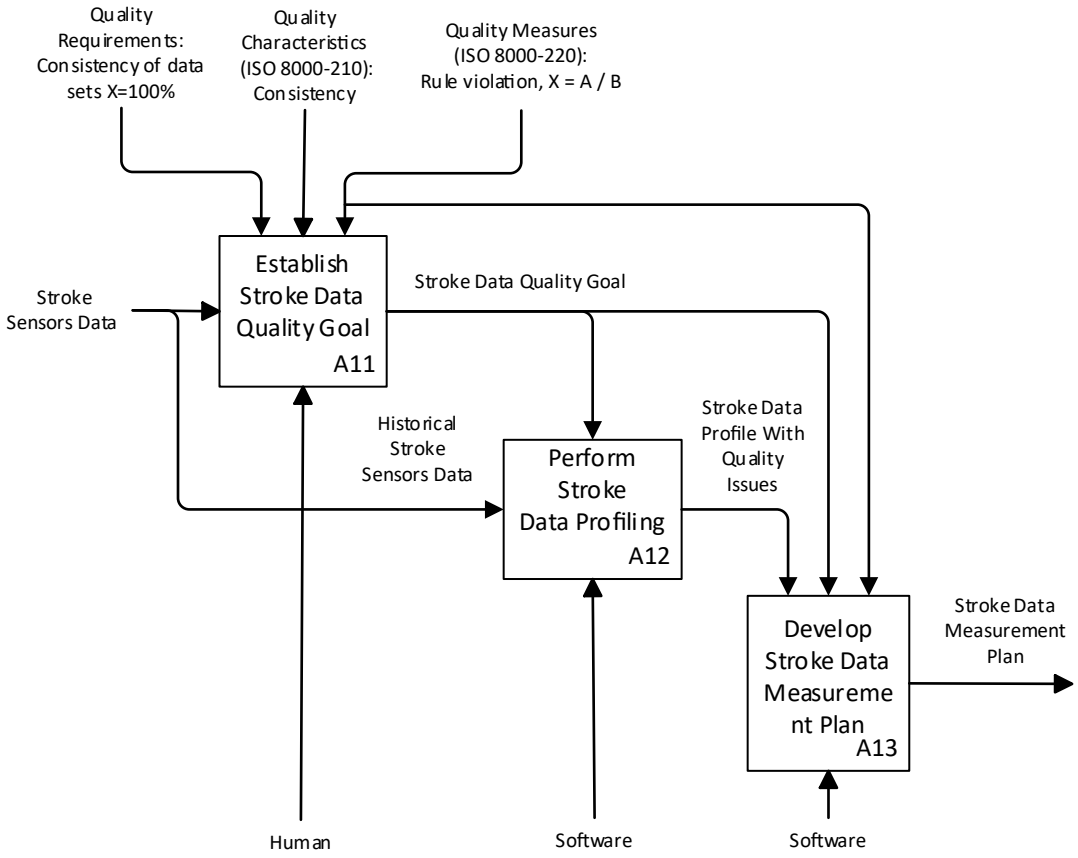


그림 D-18: 측정 계획 수립 (모델 다이어그램 A1)

D.4.3 스트로크 데이터 품질 측정 (A2)

그림 D-19에서 보이는 바와 같이, 스트로크 데이터 품질 측정 (Measure Stroke Data Quality, A2) 은 다음 두 가지 하위 활동으로 구성된다:

1) A21
   - 측정 계획에 따라, 활동 A12에서 언급된 조향 실린더의 스트로크 판정 방법을 활용하여 규칙 위반 데이터를 탐지하기 위한 이상 탐지 모델을 도출하고 스트로크 센서 데이터를 탐지(Derive Anomaly Detection Model)
   - 조향 반경에 따라 좌·우 스트로크 센서 간 허용 차이 범위(ΔL)를 설정
   - 만약 두 센서 데이터 세트 간의 차이가 이 허용 범위를 벗어나면 규칙 위반 이상을 탐지
2) A22
   - 데이터 세트의 품질 측정 일관성이 품질 요구사항을 충족한다면 추가적인 데이터 정제는 불필요

- 그렇지 않은 경우, 이상 탐지 모델을 통해 정제가 필요한 두 스트로크 센서 데이터 세트의 규칙 위반을 탐지
- 데이터 세트의 일관성은 X = 97.2%이므로 센서 데이터는 품질 요구사항 X = 100%를 충족하지 못한다. 따라서 원시 센서 데이터 세트에서 규칙 위반을 탐지하고, 규칙 위반으로 인한 품질 측정 불일치도를 계산한다. 규칙 위반으로 인한 불일치도는 Xr=2.8% 이며, 이 규칙 위반은 개선 가능하다. 따라서 품질 개선 기회가 있는 센서 데이터 세트는 다음 활동인 스트로크 데이터 품질 개선 (A3)으로 진행된다(Find Stroke Sensor Data Quality Improvement Opportunity).

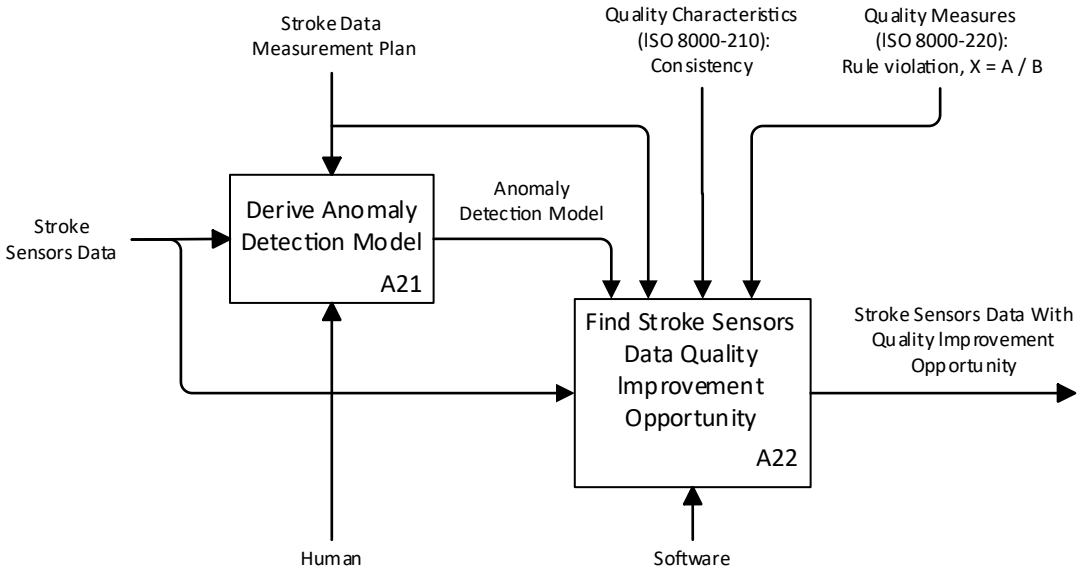


그림 D-19: 스트로크 데이터 품질 측정 (모델 다이어그램 A2)

D.4.4 스트로크 데이터 품질 개선 (A3)

그림 D-20에서 보이는 바와 같이, 스트로크 데이터 품질 개선 ((Improve Stroke Data Quality, A3) 은 다음 세 가지 하위 활동으로 구성되어 센서 데이터의 품질을 향상시킨다:

1) A31
   - 정제가 필요한 스트로크 센서 데이터 세트에 대해 데이터 보정 계획을 수립(Establish Stroke Data Repair Plan)
   - 이 방법에서는 규칙 위반의 이상값 값을 좌측 실린더의 스트로크 값 + 2초 슬라이딩 시간 창 내 좌·우 스트로크 차이의 평균값으로 대체

2) A32
   - 스트로크 센서 데이터의 책임자로부터 보정 계획에 대한 동의를 얻음(Confirm Stroke Data Repair Plan).

3) A33
   - 보정 계획이 실행 가능하다고 확정되면, 데이터 보정을 수행(Execute Stroke Data Repair)하고 정제된 스트로크 센서 데이터를 생성.

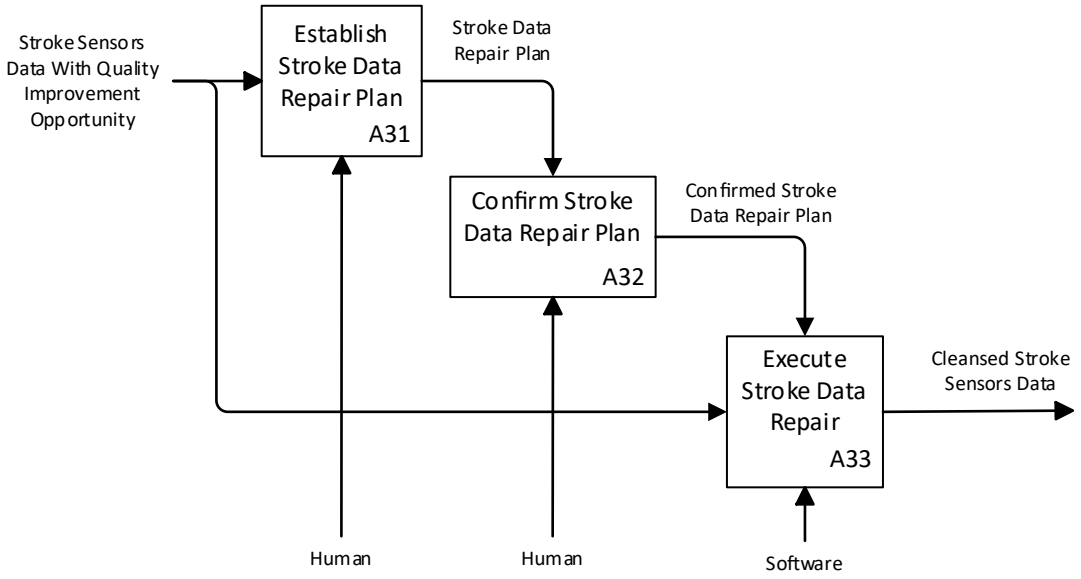
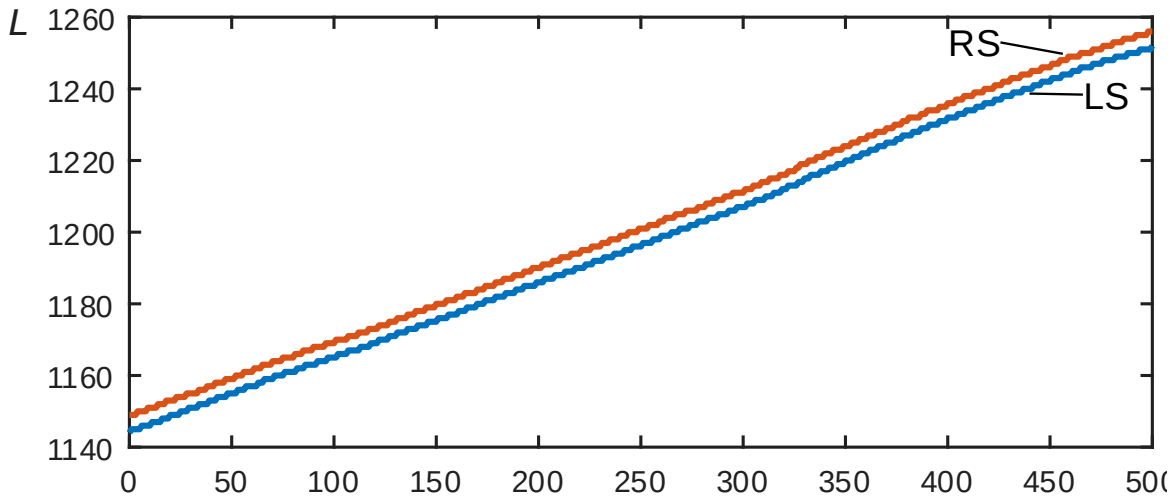


그림 D-20: 스트로크 데이터 품질 개선 (모델 다이어그램 A3)

규칙 위반 이상값을 보정한 후, 품질 개선 기회 탐색 (A22)를 다시 수행하여 정제된 스트로크 데이터 세트가 품질 요구사항을 충족하는지 확인한다. 데이터 세트의 일관성 품질 측정값이 100%로 향상되었으므로, 정제된 스트로크 데이터 세트는 품질 요구사항을 만족한다. 따라서 품질 결과 보고 (A23)를 수행하여, 품질 보고서를 포함한 품질 개선 스트로크 데이터 세트가 결과물로 생성된다.

정제 후 얻어진 품질 개선 스트로크 데이터는 그림 D-21에 나타나 있다.



Key

L 추진 실린더 스트로크, 밀리미터(mm) 단위로 표시
t 데이터 수집 시간, 초(second) 단위로 표시
LS 좌측 실린더 스트로크
RS 우측 실린더 스트로크

그림 D-21: 데이터 정제 후의 스트로크 센서 데이터