FULVIO YESID VIVAS CANTERO

SISTEMA DE CONTROL DE CALIDAD DE DATOS AGROCLIMATOLÓGICOS PARA AGRICULTURA DE PRECISIÓN

Tesis presentada a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para la obtención del Título de

Magíster en:
Ingeniería Telemática

Director:
PhD. Ing. Juan Carlos Corrales Muñoz.

Popayán
2016

Aceptación

Dedico este logro y esfuerzo a:

Mi esposa Claudia Sirley por el inmenso amor y ser mi compañera en este camino.

A mis hijos Gilber y Matthew por ser el milagro de Dios más importante en mi vida, durante este recorrido.

A mi madre Leonor y mi padre Hildebrando por su permanente apoyo y oraciones.

A mis hermanas por ese espíritu de colaboración y apoyo.

A los compañeros y amigos por dar animo en los momentos difíciles.

Agradecimientos

En este camino son muchas las personas e instituciones gracias a las cuales se ha logrado este trabajo. En primera medida debo agradecer a Dios quien ha permitido que diversos elementos se fueran dando en el momento apropiado.

A mi familia, quienes me apoyaron en todo momento.

A todos los profesores del departamento de Telemática por su apoyo constante y sus consejos en el desarrollo de la tesis.

A mi director Juan Carlos Corrales por la paciencia, insistencia y conocimientos en la estructuración de esta propuesta.

A los miembros del laboratorio de comunicaciones digitales por sus valiosos aprendizajes durante mi pasantía en Córdoba Argentina, sobre todo a Guillermo Riva y Jorge Finochieto.

Al Clúster CreaTIC por la financiación de la pasantía, por medio de los recursos asignados en la CONVOCATORIA PARA LA FORMACION DE CAPITAL HUMANO DE ALTO NIVEL PARA LA INDUSTRIA TI DEL DEPARTAMENTO DEL CAUCA 2014 - A - Capitulo 1 Maestría.

Resumen

El abastecimiento de productos agrícolas en Colombia, en su mayoría depende de los pequeños productores agrícolas, que con sus prácticas tradicionales son muy vulnerables a cambios económicos, tecnológicos, ambientales; que atentan a la seguridad alimentaria de nuestras regiones, además está en entre dicho la productividad y calidad de las cosechas de estos productores por sus grandes deficiencias y brechas entre los grandes productores donde sus productos son de tipo exportación y los rendimientos por hectárea son superiores. Ya sea para pequeños y grandes productores, hoy en día la tecnología aplicada en el los cultivos, guía la toma de decisiones estratégicas en muchas temáticas agrícolas, como es: la prevención de plagas y enfermedades, la planificación agrícola, la asistencia técnica personalizada, la elección de los cultivos, la adaptación de las prácticas agrícolas a cambios climáticos adversos.

Todo esto tiene un común denominador, la información o los datos de la parcela agrícola, como es: el suelo, clima y cultivo agrícola, es por ello que la captura de información relevante de ciertos parámetros ambientales y físicos que logren modelar una situación agrícola, es de suma importancia y reviste un especial interés en tecnologías claves para el campo como es la agricultura de precisión, que en sus etapas involucra la recolección de datos, procesamiento de información y toma de decisiones.

Después de una extensa revisión de la literatura, se observa que el control de calidad de los datos es un proceso muy importante para la agricultura de precisión que puede ser considerado en la recolección de datos. En consecuencia esta propuesta permite definir una serie de mecanismos de control de calidad de datos, que son un componente principal en la arquitectura para un sistema de adquisición de datos SAD, además como se realiza en el sitio de observación se debe tener en cuenta la información que se genera en los sensores que toman datos de fenómenos ambientales y climatológicos que son muy relevantes para un sistema de control de

calidad y es el conocimiento del ambiente que lo rodea. Este enfoque puede proporcionar a los SAD la capacidad de comprender las situaciones de su entorno con el fin de mejorar la calidad de datos para la toma de decisiones.

Como resultado de esta propuesta se tiene un arquitectura validada en un ambiente real que proporción datos con un grado de calidad aceptable, haciendo uso de mecanismos de control de calidad definidos y evaluados en un ambiente simulado para después ser un componente principal dentro del motor de calidad de datos, además la inserción de información contextual a nivel del fabricante, sitio de observación y datos generados de los sensores. Todo esto sumado a ambiente de aprendizaje por reforzamiento realizado al interior del sistema de adquisición para convertir el proceso de control de calidad en un ciclo de mejora continua.

Como conclusión se definieron una serie de mecanismos de control de calidad identificados, evaluados e implementados para ser accesible por todo el público. Incluyendo la información contextual modelada por estándares abiertos según la iniciativa SWE, además de su implementación y aporte al proceso de control de calidad en la selección y definición de parámetros, por ser un ambiente cambiante se permite generar un aprendizaje continuo que repercute en generar cada vez más datos de mayor calidad, el método usado es un aprendizaje por reforzamiento que se ejecuta localmente. La conclusión más importante es la definición de una arquitectura de referencia que incluye un motor de calidad, un motor de aprendizaje y un motor de conocimiento global y local que permita mejorar la calidad de los datos capturados por un sistema de adquisición SAD.

Palabras claves: Control de calidad de los datos, agricultura de precisión, metadatos, sistemas de adquisición de datos, modelo contextual, aprendizaje por refuerzo, SWE.

Abstract

The supply of agricultural products in Colombia depends on small farmers, with their traditional practices are highly vulnerable to economic, technological, environmental changes; that threatens food security of our regions, and it is in between that the productivity and quality of crops such producers by its major deficiencies and gaps between the major producers which export their products are kind and yields per hectare are higher. Whether for small and large producers, nowadays the technology involved in the cultivation, guide the strategic decision making in many agricultural issues, such as: the prevention of pests and diseases, agricultural planning, personalized technical assistance, choice of crops, adapting agricultural practices to adverse climatic changes.

All this has a common denominator, the information or data of the agricultural parcel, such as: soil, climate and agricultural crop, which is why the capture of relevant information from certain environmental and physical modeling parameters to achieve an agricultural situation, it is of most importance and particular interest in key technologies for the field as precision agriculture, which involves stages of data collection, information processing and decision making.

After an extensive review of the literature, it appears that the control of data quality is an important precision agriculture can be considered in the data collection process. Consequently the proposal to define a series of mechanisms of quality control data, which are a major component in the architecture for a system of data acquisition SAD, as well as performed at the observing site should take into account the information It generated in the sensors that takes data from environmental and climate phenomena that are very relevant to a system of quality control and knowledge of the surrounding environment. This approach can provide the SAD the ability to understand the

situations of their environment in order to improve the quality of data for decision-making.

As a result of this proposal has a validated architecture in a real environment to share data with an acceptable level of quality, using quality control mechanisms defined and evaluated in a simulated environment and then be a major component in the engine quality data, plus the inclusion of contextual information at the manufacturer level, site observation and data generated from the sensors. All this combined with reinforcement learning environment performed within the acquisition system to convert the quality control process in a cycle of continuous improvement.

As a conclusion a number of quality control mechanisms identified, evaluated and implemented to be accessible by all public defined. In addition to the inclusion of contextual information for open standards modeled according to SWE initiative, in addition to its contribution to the process implementation and quality control in the selection and definition of parameters to be a changing environment can generate a continuous learning impact generate more data in better quality every time, the method used is a reinforcement learning running locally. The most important conclusion is the definition of a reference architecture that includes a quality motor, motor learning and motor of global and local knowledge to improve the quality of the data captured by an acquisition system SAD.

Keywords: Data quality control, precision agriculture, metadata, data acquisition systems, contextual model, SWE, reinforcement learning.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

La propuesta definida en esta tesis se refiere a proporcionar información valiosa para la toma de decisiones de los productores agrícolas en cuanto al manejo de los cultivos, la prevención de plagas y enfermedades, planeación agrícola para asistencia técnica especializada, adaptación de prácticas agrícolas ante cambios climáticos, reducción de mano de obra, para incrementar productividad y calidad de las cosechas.

Es de reconocer que Colombia en su gran mayoría es un país rural y los alimentos provienen del trabajo de pequeños productores agrícolas que no poseen muchas herramientas tecnológicas y no aplican conceptos nuevos como la agricultura de precisión, que haga frente a los cambios climáticos adversos, como el fenómeno del niño y la niña que golpean con más fuerza esta parte del planeta.

La agricultura de precisión reconoce la variabilidad del terreno en cuanto al clima, suelo y cultivos, preservando mayor eficiencia económica, respetando el ambiente y generando productos de alta calidad. Dentro de este proceso el que recobra mayor fuerza es la captura de información por parte de un sistema de adquisición que tiene conectados una serie de sensores que capturan datos del medio, éstos sistemas están expuestos a muchos cambios climáticos que pueden generar fallas, repercutiendo en el envío de información errónea a los centro de datos donde se toman las decisiones.

Es por ello que el objetivo de ésta tesis es mejorar la calidad de los datos capturados por los sensores en el sitio de adquisición. Logrando definir una serie de mecanismos de control de calidad, una arquitectura de referencia y su respectiva evaluación en un escenario de agricultura de precisión.

1.1 Planteamiento del Problema

Los sistemas productivos en Colombia se verán afectados por el cambio climático en un 80% de los cultivos y más del 60% de sus áreas cultivables o aptas para la agricultura, al mismo tiempo los pequeños productores son responsables del 40% del valor de la producción agrícola, se hace cargo del 50% del empleo agrícola y produce más del 30% de los cultivos anuales de alimentos (LAU et al. 2011).

Unido a esta situación se suma el rezago en las prácticas y tecnologías utilizadas en el sector, debido a que sigue con tendencia de homogeneizar todas las labores agrícolas desde la siembra, la aplicación de fertilizantes, el control de plagas y enfermedades y los sistemas de riego a todo el terreno y las variaciones temporales de las plantas, los suelos y los microclimas, teniendo como resultado bajo rendimiento en las cosechas, bajos ingresos económicos y una calidad un poco aceptable de los productos llevados al mercado (Ministerio et al. 2006).

De otro lado, se tienen grandes productores agrícolas cuyos resultados se ven reflejados en los productos tipo exportación, que tiene como principal característica la utilización de tecnologías y conocimiento, entre las tecnologías están: el uso de estaciones agroclimatológicas y las redes de sensores inalámbricos que monitorean variables climatológicas y agrícolas para generar información relevante que mejora la toma de decisiones en cuanto a la planificación de los cultivos, la asistencia técnica especializada, y una producción agrícola precisa e intensiva.

En cuanto al conocimiento aplicado por parte de los grandes productores se tiene la agricultura de precisión en donde cada parte del suelo es tratado en forma única, por medio de muestras y sistemas de georreferenciación y con ayuda de la tecnología aumentar la competitividad y posibilitar un adecuado manejo ambiental de los cultivos.

La agricultura de precisión, implica considerar al menos tres etapas: la primera de ellas, la toma de datos a nivel intensivo de las variables de suelo, cultivo y microclima; la segunda, el procesamiento de dicha información y la tercera, la aplicación de respuestas adecuadas en cada labor del proceso de producción agrícola (Riopaila Castilla S.A. 2013).

Este trabajo se centra en la toma de datos a nivel intensivo de las variables de microclima en donde se analizan los datos capturados por el sistema de adquisición de datos – SAD agroclimatológicos. Estos datos de clima tradicionalmente han sido obtenidos a través de estaciones climatológicas, redes de sensores - WSN, etc. En estos sistemas de adquisición de datos la confiabilidad de los datos recae en la tecnología de comunicación inalámbrica utilizada y están basados en métricas de redes tradicionales, como son: el número de saltos, los retardos, la calidad de servicio (QoS), etc. Sin tener en cuenta la importancia y la calidad de los datos en el momento de su adquisición o procesamiento local (Ngai & Gunningberg 2014), ya que la captura de datos por parte de los sistemas de adquisición involucra varios aspectos, entre ellos: selección adecuada de los sitios de observación, sensores adecuados de medición en cuanto a confiabilidad y precisión, despliegue de nodos sensores en sitios representativos, rutinas de calibración y mantenimiento del sensor y transferencia de datos confiables a los centros de datos o sistemas de gestión y toma de decisiones.

En este marco los SAD no implementan procesos que permitan determinar si el dato adquirido trae la suficiente confiabilidad para ser insertado en los sistemas de análisis de información o son parte de un procesamiento automático en un servidor central.

De acuerdo con este contexto, la pregunta de investigación central de esta tesis es: ¿Cómo generar mayor calidad de los datos agroclimatológicos en los sistemas de adquisición de datos - SAD en la agricultura de precisión para pequeños productores agrícolas?

La hipótesis es: al usar información contextual y aplicarla en los mecanismos de detección de anomalías, gestión de alertas e incertidumbres en la recopilación de los datos en los SAD agroclimatológicos se puede generar mayor calidad de los datos para agricultura de precisión.

1.2 Escenario de Motivación

La motivación de esta tesis es la generación de datos confiables para la toma de decisiones a nivel agrícola ya que según estudios se ha comprobado que el 51% de los datos recolectados tiene errores y entre 3% y el 60% de los datos de cada sensor son defectuosos. Además de la importancia de la agricultura de precisión para abordar temas de productividad y competitividad de las cosechas, teniendo en cuenta la

variabilidad de las condiciones es que se encuentran la producción agrícola. La agricultura de precisión presenta tres fases: la aplicación de respuestas y toma de decisiones, el procesamiento de la información y la captura de datos en terreno por medio de sensores.

La presente propuesta está centrada precisamente en la primera fase de la agricultura de precisión, específicamente en la recolección de datos, que lo lleva a cabo el sistema de adquisición de datos – SAD, entre ellos están: las estaciones climatológicas y las redes de sensores, que monitorean continuamente variables agroclimatológicas (temperatura, humedad, precipitación, presión atmosférica, radiación solar, viento, temperatura del suelo y humedad del suelo). (Figura 1-1)

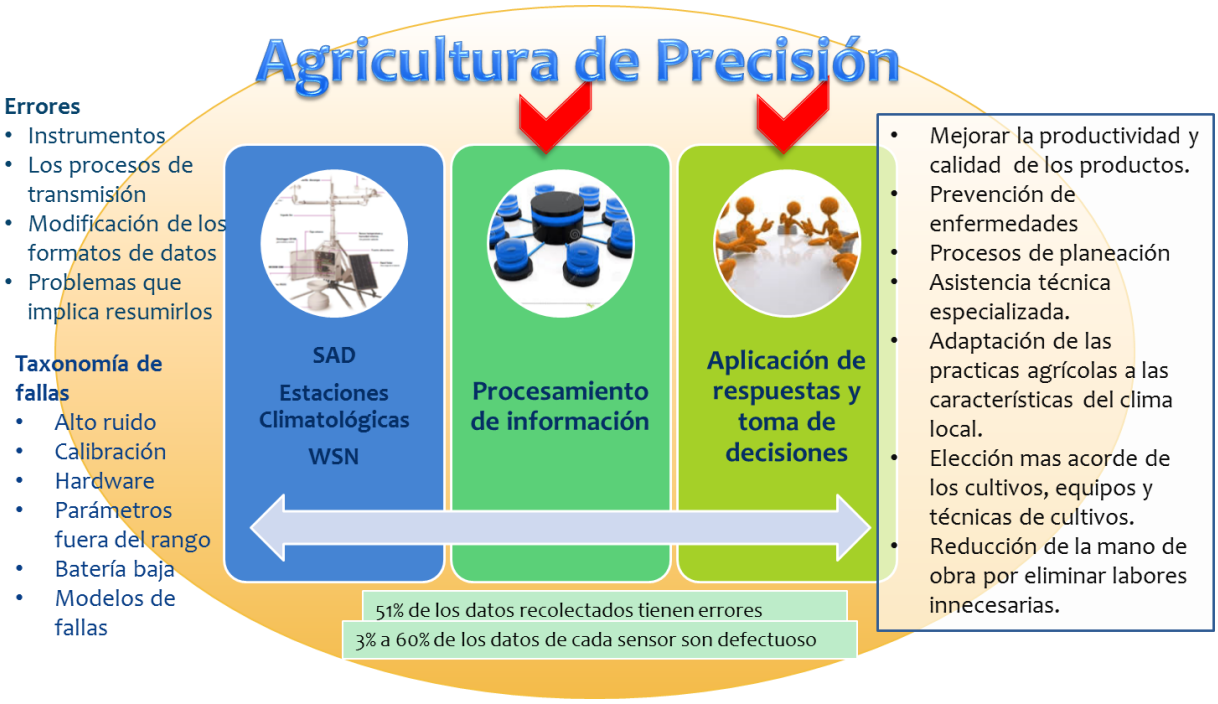


Figura 1-1 Escenario de motivación.

En esta fase de adquisición la función principal de un SAD es la medición de una variable física por medio de sensores, realizar algunas tareas simples de procesamiento y por ultimo enviar los datos a un centro de datos para su análisis. Con base en esto, la energía se consume en tres formas: adquisición, procesamiento de datos y operaciones de comunicación. Estos sistemas son desatendidos ya que se despliegan en un área determinada del cultivo, reciben poca intervención humana y son susceptibles a fallas.

La taxonomía de fallas propuesta por (Ni et al. 2009) en la etapa de adquisición del SAD se clasifican en: centrada en los datos recolectados de los sensores (parámetros fuera del rango, valores atípicos, pocas o nulas variaciones para un periodo de tiempo más largo de lo esperado y ruido excesivo) y fallas centrada en el sistema (falta de calibración de los instrumentos, fallas en el hardware, batería baja, los valores ambientales superan las capacidades del sensor), para la etapa de transmisión la confiabilidad de los datos según (Ngai & Gunningberg 2014; Ji et al. 2011) se basa en métricas de redes tradicionales.

La mitigación o la reducción de todos los tipos, fuentes y causas posibles de errores o fallas por medio de un conjunto de mecanismos y procedimientos es lo que se llama un control de calidad de los datos y es el enfoque de nuestra propuesta.

En agricultura de precisión la mayoría de procesos y mecanismos de control de calidad de los datos se ejecutan en el centro de datos (Muller et al. 2013; Hubbard et al. 2012; Gwilliams et al. 2012), sobre una base de datos y hacen uso del contexto que según Dey (Dey 2001) “Contexto es cualquier información que puede ser caracterizada para definir una situación de una entidad” y almacenados como metadatos (Muller et al. 2013; Hubbard et al. 2012). Los metadatos son el resultado de la interacción y el conocimiento de información relacionada con el ambiente y la información procesada de otros sistemas de adquisición llamada información contextual que ayuda a modelar la situación de recolección de datos.

Por ultimo nuestra propuesta se enmarca en definir y adaptar ciertos mecanismos de control de calidad de datos en el SAD, planteando una arquitectura que haga frente a los recursos limitados de los sistemas de adquisición e integre los metadatos para enriquecer el proceso de control de calidad (Organizacion Meteorológica Mundial - OMM 2011).

1.3 Objetivos

1.3.1 Objetivo General

Proponer mecanismos de control de calidad de datos agroclimatológicos en los sistemas de adquisición de datos – SAD para ser aplicados en la agricultura de precisión.

1.3.2 Objetivos Específicos

- Definir mecanismos de procesamiento de información contextual que mejore la calidad de los datos recopilados del sistema de adquisición de datos - SAD en la agricultura de precisión.
- Especificar una arquitectura de referencia que soporte los mecanismos de control de calidad de los datos agroclimatológicos en el sistema de adquisición de datos - SAD.
- Desarrollar y evaluar experimentalmente un prototipo que implemente la arquitectura propuesta.

1.4 Contribuciones

Mediante este trabajo de grado se logra definir un sistema de control de calidad de datos agroclimatológicos para agricultura de precisión, conjunto de mecanismos de control de calidad de los datos aplicados a los sistemas de adquisición de datos agroclimatológicos - SAD para la agricultura de precisión.

- ✓ Un conjunto de mecanismos de procesamiento de datos basados en información contextual entre ellos: comprensión, calidad, generación de nueva información, configuración y estandarización.
- ✓ Una arquitectura de referencia basada en estándares de la OGC (Sensor Web Enablement - SWE) e información contextual para el sistema de adquisición de datos agroclimatológicos.
- ✓ Un prototipo experimental del sistema de adquisición de datos agroclimatológicos.

1.5 Contenido de la Monografía

Capítulo 2. Estado actual del conocimiento

En el desarrollo de la propuesta de tesis primero se presenta unos conceptos generales sobre agricultura de precisión, sistemas de adquisición de datos, anomalías y mecanismos de control de calidad, en la segunda parte se mencionan algunos trabajos relacionados u otras investigaciones que se han realizado en el proceso de control de calidad de los datos y detección de anomalías; en la tercera parte el análisis de brechas.

Capítulo 3. Mecanismos de procesamiento de la información

La propuesta presentada en este trabajo, se refiere a los sistemas de adquisición de datos presentes en una aplicación de agricultura de precisión y dentro de ella se abordara la etapa de adquisición y procesamiento de los datos, teniendo en cuenta las diferentes fallas y anomalías presentes en estos sistemas por ser limitados en recursos, el estudio se centra en un control de calidad de los datos que involucra la detección de anomalías a nivel de datos: parámetros fuera de rango, valores atípicos, fallas constantes y ruido excesivo. En este capítulo adicionalmente se presenta el escenario de control de calidad de datos para agricultura de precisión, los mecanismos propuestos y la evaluación de los mismos en un ambiente de simulación (OMNET) para validar el grado de detección de anomalías.

Capítulo 4. Arquitectura de referencia

En este capítulo se describe la arquitectura del sistema propuesto, teniendo en cuenta la detección de anomalías, la información contextual de fuentes externas y la información contextual generada en el propio sistema para actualización de parámetros de referencia y que el proceso de control de calidad sea un ciclo de mejora continuo. Al inicio se definirán algunos modelos que están implícitos en un sistema de adquisición y otros que permitan cumplir con las funcionalidades de un control de calidad, enseguida se muestran los componentes de la capa de control de calidad, su modelo lógico y por ultimo sus diferentes niveles de abstracción en donde están inmersos los mecanismos de control de calidad definidos en el capítulo 3.

Capítulo 5. Evaluación de la propuesta

En este capítulo se implementa la arquitectura del sistema propuesto junto con los mecanismos de calidad de datos seleccionados, la cual es objeto de evaluación en un ambiente real. Se selecciona la plataforma embebida Waspmote Plug & Sense y la tarjeta de agricultura inteligente (Smart agricultura PRO) obteniendo de ella información para el modelamiento del sensor. Igualmente se selecciona el sitio de observación y se recopila la información más relevante para modelar un contexto ambiental.

Al inicio se modela cada uno de los componentes de la arquitectura propuesta (la información contextual y los mecanismos de control de calidad), se crea una librería

para ser adaptable al entorno de desarrollo, se construye la lógica del programa principal con base en los niveles de control de calidad definidos en el capítulo 4 y se evalúa todo el sistema por los escenarios descritos en el capítulo 3, etiquetando los datos de salidas con marcas de calidad.

Capítulo 6. Conclusiones y trabajos futuros

Proporciona las conclusiones de este trabajo, así como una discusión sobre los posibles trabajos futuros que se derivan de esta propuesta de tesis o de algunos apartados que están por fuera del alcance del sistema de control de calidad de los datos agroclimatológicos para agricultura de precisión.

Capítulo 2

Estado actual del conocimiento

En el desarrollo de esta propuesta, primero se presentan unos conceptos generales sobre agricultura de precisión, sistemas de adquisición de datos, anomalías y mecanismos de control de calidad, en la segunda parte se mencionan algunos trabajos relacionados u otras investigaciones que se han realizado en el proceso de control de calidad de los datos y detección de anomalías; en la tercera parte el análisis de brechas.

2.1 Agricultura de precisión

Es un concepto agronómico de gestión de parcelas agrícolas, basado en la existencia de variabilidad en campo. Requiere el uso de las tecnologías (Atzberger 2013) de Sistemas de Posicionamiento Global (GPS), sensores, satélites e imágenes aéreas junto con Sistemas de Información Geográfica (SIG) para estimar, evaluar y entender dichas variaciones (Lee et al. 2010). La información recolectada puede ser usada para evaluar con mayor precisión la densidad óptima de siembra, estimar fertilizantes y para predecir con más exactitud la producción de los cultivos.

La agricultura de precisión tiene como objeto optimizar la gestión de una parcela desde los siguientes puntos de vista.

- ✓ Agronómico: ajuste de las prácticas de cultivo a las necesidades de la planta.
- ✓ Medioambiental: reducción del impacto vinculado a la actividad agrícola.
- ✓ Económico: aumento de la competitividad a través de una mayor eficacia de las prácticas agrícolas.

Además, la agricultura de precisión se basa en el reconocimiento de la variabilidad espacio temporal intrínseca que supone el manejo de un cultivo, considerando tres etapas: i) la recolección de datos a nivel intensivo de las variables de suelo, cultivo y microclima; ii) la generación de información y iii) el uso de esta información en la toma de decisiones (Riopaila Castilla S.A. 2013) (Peets et al. 2012).

Unos estudios sobre dedicación de tiempo en la gestión de información realizados por (Gasparin 2009) sobre la implementación de un sistema automático para gestión de parcelas agrícolas da cuenta que un 49% de esfuerzo implica la recopilación de la información, un 16% en el análisis de datos y un 35% en la toma de decisiones. Por tanto, la fase de recolección de datos es una fase crítica en agricultura de precisión.

La propuesta está centrada precisamente en la primera fase de la agricultura de precisión, específicamente en la recolección de datos, que lo lleva a cabo el sistema de adquisición de datos – SAD.

2.2 Sistemas de adquisición de datos - SAD

Es un sistema que busca capturar la información de un parámetro físico, ambiental, químico de una zona de estudio, para ser enviada a un centro de datos donde es analizada e interpretada para apoyar la toma de decisiones; los componentes que integran el sistema van desde el dispositivo sensor, el circuito acondicionador de señal, el conversor análogo digital, el microcontrolador que procesa dichos datos, un almacenamiento temporal, el circuito de transmisión y la fuente de energía que alimenta todo el sistema de adquisición.

Para agricultura de precisión, los SAD son utilizados en la etapa de recolección de datos, entre los más utilizados están: las estaciones climatológicas y las redes de sensores inalámbricos – WSN.

- ✓ Una estación climatológica es una instalación destinada a medir y registrar regularmente diversas variables meteorológicas a nivel de meso escala. Encargada de registrar mediciones ordenadas en el tiempo, provenientes de diferentes sensores (Temperatura, Humedad, Presión atmosférica, Viento, Precipitación, Radiación, etc.).
- ✓ Las redes de sensores inalámbricos (WSN) integrada por diferentes sensores, nodos sensores y tecnologías de comunicación, que capturan datos de un

fenómeno (humedad, temperatura, etc.) y los envía a un servidor central para ser procesados (Akyildiz et al. 2002). La arquitectura de la WSN está centrada en los datos, sus recursos son limitados en cuanto a procesamiento, energía y almacenamiento (Wang & Balasingham 2010).

Los SAD tienen la capacidad de detectar su entorno en donde se realiza la medición de una variable física por medio de sensores, procesar la información a nivel local, almacenamiento temporal de dicha información y enviar a uno o más destinos la información por lo general va al centro de datos para su posterior análisis.(Konieczny 2012). (Figura 2-1).

Figura 2-1 Etapas de un sistema de adquisición de datos.

El diseño de un SAD presenta una serie de desafíos entre los más frecuentes están:

a) Los sistemas de adquisición para agricultura de precisión incluyen nodos sensores heterogéneos, distribuidos, con alto grado de autonomía y requisitos de auto configuración.
b) En el escenario de monitoreo ambiental en donde hacen uso de las WSN, el desafío más frecuente es el factor de consumo energético en un nodo debido al sistema de comunicación inalámbrica (Oliveira & Rodrigues 2011). A mayor comunicación, menor vida útil. Se estima que transmitir o recibir 1 bit a 100 metros de distancia requiere la misma cantidad de energía (ETx ERx EIdle = 50nJ=bit) que la necesaria para computar 3000 instrucciones de código con un microcontrolador de 8 bits de bajo consumo (Asada et al. 1998; Bult et al. 1996).
c) Los datos generados por los sensores son continuos y periódicos, correlacionados temporal y espacialmente, y muy ruidosos debido al ambiente donde son desplegados; presentando una serie de errores, fallas y anomalías que afectan la confiabilidad de los datos.

2.3 Anomalías (métricas y detección)

En los sistemas de adquisición de datos el problema es que los datos capturados por los sensores son imprecisos e incompletos, la imprecisión se debe a que los nodos son susceptibles a presentar fallas por la limitación de recursos (Akyildiz et al. 2002)(memoria, procesamiento, energía y ancho de banda) y se evidencia en lecturas erróneas del fenómeno a ser monitoreado. Otro tipo de imprecisiones están relacionadas con el ambiente donde los nodos sensores son desplegados ya que son sitios remotos, con condiciones adversas, y desatendidos en su operación y gestión.

Sumado a esto, el despliegue de cientos y miles de nodos sensores en un área de gran tamaño enviando muchas veces datos redundantes generando imprecisión en la información. Además, los datos incompletos en el destino es el resultado de comunicaciones poco fiables debido a la forma como los nodos pasan de un estado inactivo a otro activo, cambiando dinámicamente la red.

Las mediciones de datos inexactos, imprecisos o incompletos causados por las razones antes mencionadas se conocen como anomalías (Van Essen et al. 2012). Una anomalía se define en (Rassam et al. 2013), una observación que parece ser incompatible con el resto de un conjunto de datos o patrones en los datos que no se ajustan a un comportamiento normal bien definido. Otros términos relacionados con anomalías son valores atípicos, fallas y desviaciones.

Las anomalías en sistemas de adquisición se clasifican (Jurdak et al. 2011) en: i) anomalías en la red , ii) anomalías en el nodo y iii) anomalías en los datos.

- ✓ Anomalías de la red son problemas relacionados con la comunicación que surgen en la WSN. Sus típicos síntomas están relacionados con un aumento inesperado o disminución de la cantidad de paquetes que atraviesan la red. Causados por mala conectividad, conectividad intermitente, existencia de ciclos de enrutamiento, tormentas de *broadcast*. El alcance de estas anomalías es de muchas instancias de datos o todo el conjunto de datos y la métrica de detección es la calidad del enlace, el número de paquetes enviados y recibidos, la identificación del paquete, etc.
- ✓ Anomalías a nivel del nodo están relacionadas con problemas de hardware o software en un solo nodo y no están relacionados con la comunicación con los

nodos vecinos. Debido a que surgen en nodos individuales, la detección distribuida puede ser muy efectiva. Un síntoma común a nivel de nodo es cuando un nodo detiene la transmisión de datos. La causa más probable de esta anomalía es la falla / degradación de los paneles solares, lo que conduce a perdida de alimentación en la tarjeta. Otras fallas son reinicio del nodo, problemas en la batería y fallas en el nodo. El alcance es a nivel del nodo y las métricas para detección se basan en verificar el nivel de la fuente de alimentación (panel solar, batería), el contador de paquetes con un valor igual a cero, etc.

- ✓ Anomalías de datos depende de irregularidades estadísticas en los datos. Estas irregularidades pueden ser causadas por el hardware del sensor, mala calibración, sensores defectuosos o variaciones ambientales. Las fallas del sensor reportan valores extremos o no realistas, las anomalías de datos presentan tres categorías: 1) las anomalías temporales que existe un cambio en los valores de los datos capturados por el sensor sobre el tiempo y se detectan localmente, 2) las anomalías espaciales donde se compara el valor del dato de un nodo con sus vecinos y 3) las anomalías espaciotemporales se detecta a través de un numero de nodos donde hay cambios en los valores de los datos sobre el tiempo y el espacio. Estas últimas se detectan por medio de la interacción con otros nodos (Jurdak et al. 2011).

Otra clasificación de tipos de anomalías las menciona (Ni et al. 2009) (Lee et al. 2010) presentada como una taxonomía de fallas: i) fallas centrada en los datos recolectados de los sensores (parámetros fuera del rango, valores atípicos, pocas o nulas variaciones para un periodo de tiempo más largo de lo esperado y ruido excesivo) y ii) fallas centrada en el sistema (falta de calibración de los instrumentos, fallas en el hardware, batería baja que influye en el funcionamiento del sensor, los valores ambientales superan las capacidades del sensor), estas fallas generan errores en los datos y afectan la calidad de los mismos.

En resumen las anomalías que se tienen en cuenta en este estudio son anomalías de datos (Ravichandran & Arulappan 2013) y entre ellas las siguientes:

- Parámetros fuera del rango - *Out-of-range*: son muestras de datos de sensores que se desvían significativamente de rango esperado de valores. Representan valores de sensores que físicamente no son posibles en la región de estudio.

- Valores atípicos - *Outliers*: son datos aislados que se desvían significativamente de las otras instancias de la muestra, pero aparecen dentro del rango esperado de valores.
- Fallas constantes - *Struck-at*: Pocas o nulas variaciones para un periodo de tiempo más largo de lo esperado, los datos se congelan o permanecen en un valor dado, puede estar dentro o fuera del rango esperado.
- Ruido excesivo – *Spike*: La tasa de cambio en el gradiente de las muestras de datos sobre el período de tiempo es mucho mayor de lo esperado. Se produce en combinación de al menos algunas muestras de datos sucesivos.

2.4 Control de calidad de los datos

La existencia de anomalías, fallas y errores, repercute en la toma de decisión. Ya que los análisis son producto de datos no confiables, es por ello que se busca diseñar modelos eficientes y efectivos en la detección de anomalías, esto es lo que se llama un control de calidad de los datos.

La detección de anomalías es el proceso que identifica un patrón de datos que se desvía del comportamiento esperado. Es diseñar un modelo que detecte un comportamiento anormal en el flujo de datos de los sensores.

Los mecanismos de control de calidad de los datos mitigan o reducen todos los tipos, fuentes y causas posibles de errores; mejorando la calidad de los datos capturados, optimizando la robustez del análisis de datos en virtud de la presencia de ruido, evitando que el resultado de la agregación de datos pueda ser afectada.

Dichos correctivos van desde la calibración del sensor, la no inserción de los datos erróneos en el centro de datos, la programación de tareas de mantenimiento de los sensores, la manipulación de otros mecanismos que den cuenta de los posibles valores a ser analizados.

En conclusión la importancia de un control de calidad de los datos radica en proporcionar datos confiables y de alta calidad siendo la tarea más crítica e importante en los procesos de análisis, optimización y toma de decisión al nivel del negocio. (Figura 2-2).

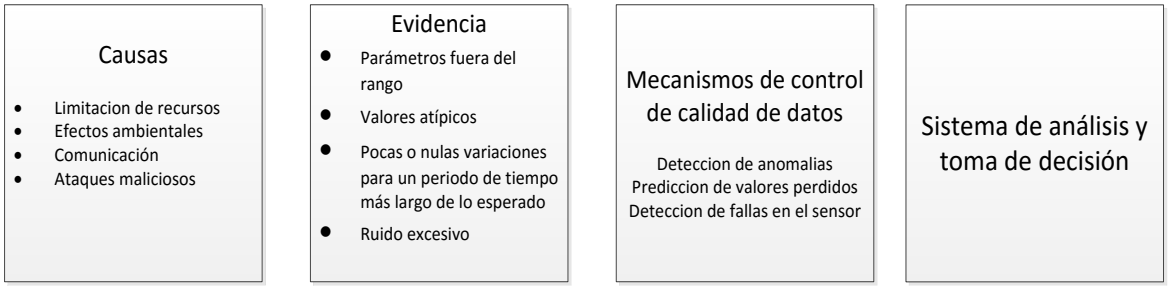


Figura 2-2 Ambiente de control de calidad de datos.

Los mecanismos de control de calidad de los datos agroclimatológicos y sobre todo en agricultura de precisión, ayudan a: mejorar la productividad y calidad de los productos, prevención de enfermedades, optimizar procesos de planeación, asistencia técnica especializada, adaptación de las prácticas agrícolas a las características del clima local, elección más acorde de los cultivos, equipos y técnicas de cultivos, reducción de la mano de obra por eliminar labores innecesarias.

Para la Organización Mundial de Meteorología OMM el control de calidad de los datos es un área de estudio muy importante y con avances significativos, dentro de sus guías (Organizacion Meteorológica Mundial - OMM 2011) se encuentra un modelo que deben cumplir las estaciones climatológicas. (Figura 2-3).

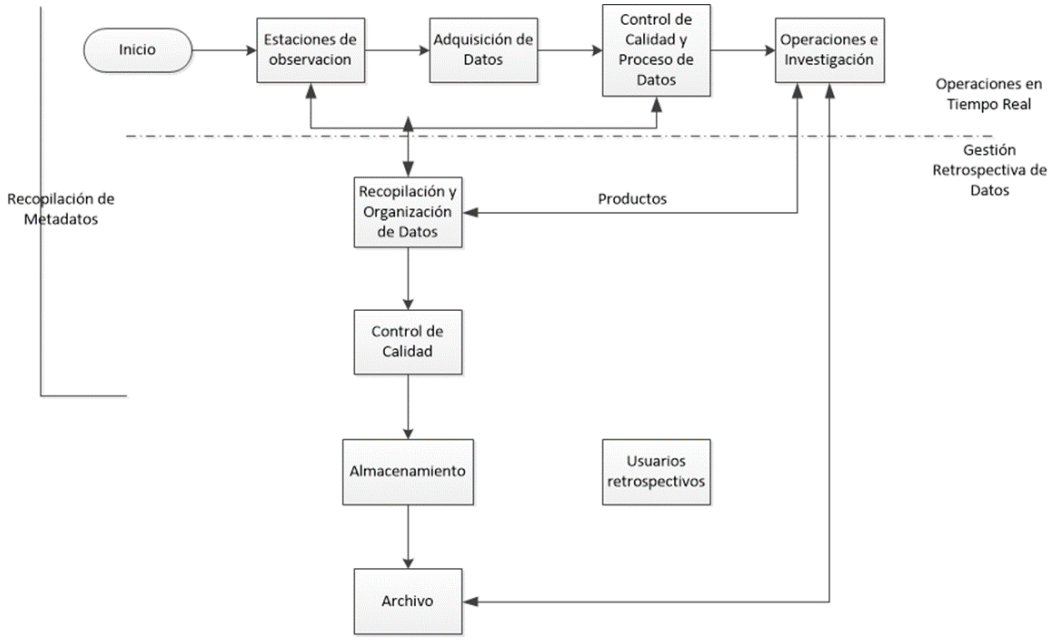


Figura 2-3 Modelo de control de calidad de datos según las OMM.

2.5 Mecanismos de control de calidad de los datos

Existen muchas alternativas de control de calidad de datos entre ellas: técnicas de detección livianas por ser compactas y eficientes, reconstrucción de esquemas de detección distribuidas por toda la red y reduciendo el encabezado de los datos cuando se envían a otros destinos, patrones de detección que conservan la energía sin perder la confiabilidad, otras estrategias están ligadas a comprimir los datos, limitar el número de atributos, y simplificar el procedimiento de análisis y decisión.

Los esquemas de detección de redes tradicionales no son aplicables a los sistemas de adquisición entre ellos la WSN (Tseng et al. 2013) y el desafío clave es la efectividad en la detección de anomalías y eficiencia en la utilización de recursos sobre todo minimizando el costo de energía, para prolongar el tiempo de vida de la red (Meratnia & Havinga 2010) y en su gran mayoría aplicados sobre la base de datos en un servidor central.

Algunas técnicas implementadas en los sistemas de adquisición son innovadores en su aplicación entre ellas tenemos: basadas en clasificación que implica un clasificador, datos de entrenamiento y pruebas (redes neuronales, redes bayesianas, máquinas de vector de soporte - SVM, basadas en reglas); basadas en el vecino más cercano (distancia K *nearest neighbor*, densidad relativa), basadas en *clustering*, basadas en pruebas estadísticas (técnicas paramétricas como los modelos *gaussianos* y de regresión y las no paramétricas como los histogramas).

Cabe resaltar que la aplicación de cada técnica depende de ciertos criterios de selección (chandola, 2009) como son: disponibilidad de datos de entrenamiento, datos etiquetados o datos brutos, tipo de anomalías, tipos de entrada de datos, tipos de salida de datos, naturaleza de los datos, limitaciones y requerimientos del sistema. (Figura 2-4).

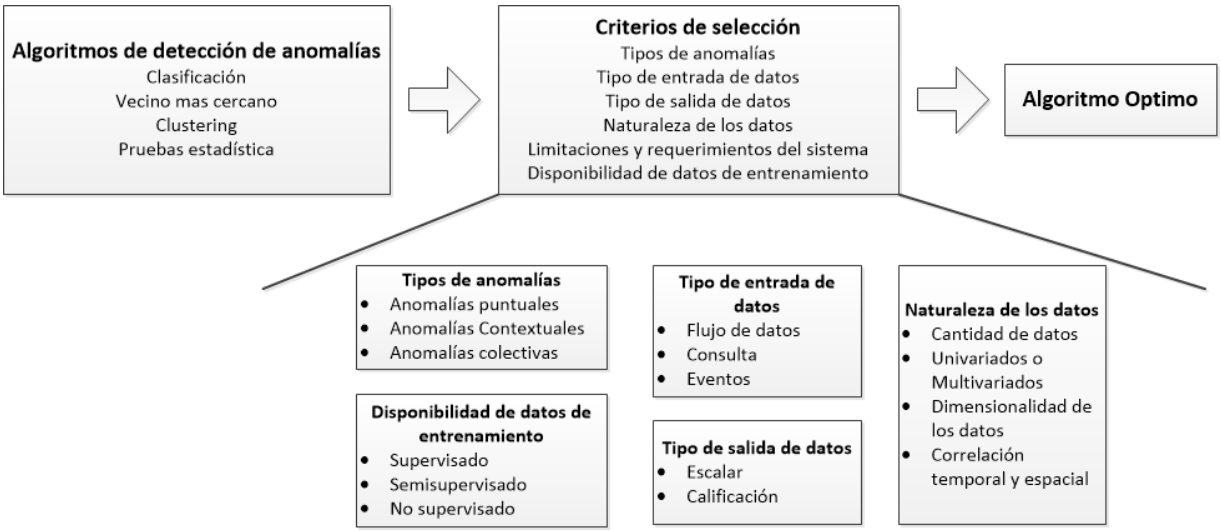


Figura 2-4 Criterios de selección de mecanismos para detección de anomalías.

Tipos de anomalías

- ✓ Anomalías puntuales: Una instancia de datos individual se puede considerar como anómala con respecto al resto de los datos; cuando los valores de ciertas instancias O_1, O_2 y O_3 se encuentran fuera de los límites de un grupo N_1 o N_2. Este tipo de anomalía es las más usada por los investigadores en las técnicas de detección.
- ✓ Anomalías contextuales: Una instancia de datos es anómala en un contexto específico, pero no en otro. Tal contexto debe ser especificado como parte de la formulación del problema y cada instancia de datos es definida con atributos: contextuales y comportamentales. Por ejemplo, si se observa las instancias de datos en una serie de tiempo muy pequeña puede marcarse como valores anómalos pero si la serie de tiempo es mayor nos puede ofrecer un evento que es normal.
- ✓ Anomalías colectivas: Un conjunto de instancias de datos es anómala con respecto a todo el conjunto de datos. Ya que son indistinguibles de los datos esperados, y sólo se pueden identificar teniendo en cuenta todo el conjunto de datos.

Entrada de datos

Existen tres modelos de entrega de datos desde el nodo sensor al centro de datos, entre ellos está por eventos, por consultas y de forma continua. (Liu 2014)

- ✓ Consulta, donde el patrón de comunicación es solicitud – respuesta, donde el nodo central decide y controla cuando recopilar datos, haciendo una solicitud a un nodo sensor y este responde pasivamente al comando con los datos capturados por los sensores.
- ✓ Forma continua o flujo de datos, en este enfoque los nodos sensores capturan los datos y deciden cuando enviarlos al nodo central de forma periódica, el nodo central tan solo almacena y procesa las ráfagas de datos recibidas.
- ✓ Eventos, los nodos centrales y los nodos sensores toman el control de cuando transmitir y recibir datos, basados en un modelo de eventos que establece ciertas responsabilidades, en los nodos centrales se establecen criterios de búsqueda del dato sensado y así recuperar la información haciendo una consulta. Y los nodos sensores deciden cuando enviar con base en umbrales o condiciones establecidas para el dato sensado.

Disponibilidad de datos de entrenamiento o datos iniciales de un conocimiento anterior

Se refiere a la existencia o no de información clasificada y etiquetada como anómala o normal, que pueda ser usada en la detección de anomalías, ya sea como una base de conocimiento inicial o para datos de entrenamiento de alguna técnica de detección; por lo general las técnicas más comunes son:

- ✓ Los algoritmos de aprendizaje supervisado, donde se hace uso de datos etiquetados y clasificados como normales y anómalos. En este aprendizaje se tiene un entrenador que presenta un patrón de datos de entrenamiento y su correspondiente objetivo de salida generando un modelo predictivo, las instancias de datos son comparadas con el modelo para determinar a qué clase pertenece (Kulkarni et al. 2011).
- ✓ El algoritmo de aprendizaje no supervisado, el objetivo es descubrir patrones en los datos de entrada sin la ayuda de un entrenador. No existe alguna evidencia de clasificación de los datos; además, no se puede inferir alguna marca o etiqueta de los datos como anómala u otra similar.
- ✓ Los algoritmos de aprendizaje semisupervisado, involucran técnicas que asumen solo datos de entrenamiento de instancias etiquetadas como datos normales, creando un modelo de comportamiento normal y usa este modelo para identificar anomalías.

Salida de datos

Es importante determinar cómo se reporta la anomalía, después de aplicar las técnicas de detección; existen dos tipos: a) la calificación en donde cada anomalía tiene una puntuación presentando un ranking de anomalías y el resultado de la detección es función del grado en que dicha instancia de datos se considera una anomalía, b) etiquetada, se asigna una etiqueta normal o anómala a una instancia de datos.

Características de los datos

Datos de sensores se recolectan en forma de flujos de datos que pueden ser grandes volúmenes de muestras obtenidas del medio ambiente descritas por medio de un atributo o característica. Algunos sistemas de adquisición están diseñados sólo para recopilar un atributo llamado univariados, como la temperatura, la luz y la humedad. Los recientes sistemas de adquisición están diseñados para recopilar múltiples atributos, ya sea del mismo tipo o tipos de datos diferentes que se llaman datos multivariados.

Una medición de los sensores es anómala si uno o más de sus atributos son anómalos. Con los datos univariados, la detección de anomalías se observa en el atributo de datos y se compara con otras instancias de datos. Aunque el análisis de datos multivariados es alto computacionalmente, la detección de anomalías es de gran precisión, si las relaciones entre los diferentes atributos se seleccionan cuidadosamente.

Sumado a esto los datos de los sensores presentan correlación espacial y temporal entre las diferentes mediciones del sensor. La correlación temporal significa que las muestras recopiladas en una serie de tiempo se relacionan entre sí. Las correlaciones espaciales significan que los datos de nodos cercanos geográficamente son muy similares.

Además, al aumentar el número de nodos desplegados en la red, la dimensionalidad de los datos capturados también aumenta; incurriendo en un alto costo computacional en cuanto a consumo de energía y memoria, Y también se incrementa el costo de transmisión por enviar mayor número de instancias desde el nodo sensor a la estación base.

Modo de operación

Los modelos de detección de anomalías se realizan después de una ventana de tiempo especificado por el diseño de la aplicación. Aunque los modelos de detección fuera de línea consumen menos energía, requieren memoria adicional para el almacenamiento de los datos en la ventana de tiempo. Además, la integridad de los datos puede verse afectada debido al retardo de tiempo en la detección, por lo tanto, la detección en línea es la preferida para minimizar el retardo de tiempo y garantizar la integridad de los datos.

Un factor adicional que afecta el uso del modo de detección en línea es el costo computacional de los métodos de detección. La eficiencia de cualquier solución de detección de anomalía no se ve afectado solamente por la dimensión de los datos, sino también por la complejidad computacional de los métodos utilizados para la detección.

Estructura del modelo

Se adoptaron tres tipos de estructuras para los modelos de detección de anomalías existentes que son locales, centralizadas, y distribuidas. En una estructura local, el detector de anomalías se implementa en el ámbito del nodo sensor sin la colaboración entre los nodos de la red. En una estructura centralizada, todos los datos se envían a una estación central, se analiza la información y converge en una base de datos donde el proceso de detección de anomalías se lleva a cabo. Por último, la estructura distribuida adopta la colaboración entre los nodos para el mecanismo de detección, en el que cada nodo envía un resumen de sus datos representados por su modelo de referencia local para enviar a la estación base y construir el modelo de referencia global, luego es utilizada por cada nodo del clúster para su posterior detección.

La estructura centralizada consume demasiada energía ya que posee más información que enviar al nodo central. Por lo tanto, la estructura distribuida es la preferida con el fin de minimizar el consumo de energía, evitando una gran cabecera de los datos del modelo de referencia local enviado al nodo central para reducir la sobrecarga de comunicación.

2.5.1 Clasificación (Redes Bayesianas, SVM, reglas)

En el modelo se tiene un clasificador que es entrenado utilizando patrones de datos conocidos y después clasifican los patrones desconocidos en uno o más tipos. Los clasificadores supervisados de múltiples clases no son utilizados en WSN por la dificultad de obtener datos etiquetados. Mientras tanto, los clasificadores supervisados de una clase son el tipo de clasificador que aprenden los patrones normales y consideran cualquier patrón fuera de lo normal como anomalía. Este clasificador es más adecuado en WSN, debido a que los patrones normales se pueden capturar y almacenar para entrenar el clasificador.

Las técnicas más utilizadas son: las redes bayesianas de *Naïve - Naive Bayesian Network* (NB) (Hill & Minsker 2010) , las redes bayesianas de *Belief - Belief Bayesian Network* (Janakiram et al. 2006) (BBN), y las redes bayesianas dinámicas - *Dynamic Bayesian Network* (DBN), el problema de aprendizaje para correlación espacial y temporal de los datos del sensor es mapeado al aprendizaje de los parámetros NB y luego realizar la inferencia. Mientras tanto, las dependencias condicionales entre atributos de datos son explotados para aumentar la precisión de la detección. La DBN permite hacer frente a los cambios dinámicos de la WSN en el tiempo.

Otras técnicas utilizadas son: las máquinas de vector de soporte, *quarter sphere* SVM (QSSVM) que se ejecuta en la estructura distribuida donde cada nodo obtiene el valor del QSSVM y lo envía a la estación central, en él se calcula un patrón global junto con su valor local de referencia, al finalizar se envía nuevamente a los nodos para detectar anomalías. Otra técnica es la *hyper-ellipsoid one-class* SVM en donde las correlaciones son consideradas, al igual que la actualización del modelo de referencia normal, el análisis de componentes principales ayudan a establecer umbrales predefinidos en los datos, clasificando como normal o anómalo (Rassam et al. 2013).

2.5.2 Vecino más cercano (*Nearest Neighbor*)

Modelos basados en el vecino más cercano con la suposición de que el patrón normal de datos siempre se encuentra en los vecinos más cercanos y las anomalías están en los vecinos más lejanos (Chandola et al. 2009). El concepto de este modelo se basa en el uso de medidas de similitud que miden el grado de un patrón de datos entre normal o anómalo, tal como la medida de la distancia euclidiana.

En este enfoque cada nodo calcula su patrón local con base en la distancia euclidiana y la envía a toda la red donde se obtiene un patrón global, que es usado para detección de anomalías; la efectividad se da en anomalías simples y de corto tiempo, para anomalías de gran duración se debe reducir la dimensionalidad de los datos por medio del análisis de componentes principales (PCA) (Xie et al. 2011); otras propuestas (Xie et al. 2011) se basan en el algoritmo K-NN pero se basan en conjunto de datos homogéneos y estáticos. Donde varias instancias de los datos se conectan en un espacio con su instancia más cercana, generando grupos, inclusive muchos más pequeños que son limitados por el algoritmo K-NN o *spanning tree*.

2.5.3 Modelos estadísticos

Los modelos de detección de anomalías basados en estadística son usados principalmente para un conjunto de datos de una dimensión. El principio esencial es construir un modelo estadístico de datos normales, en forma de distribución de probabilidad que representa la distribución de los datos en un modelo de referencia y evaluar cada patrón con respecto a ese modelo de referencia. Cualquier desviación del modelo de referencia se considera como anomalía. Muchas técnicas estadísticas se han utilizado para la detección de anomalías en WSN y categorizado en técnicas paramétricas y no paramétricas. En la categoría paramétrica, se supone que los datos se generan a partir de una distribución conocida y los parámetros de la distribución se calculan fácilmente a partir de estos datos. En la categoría no paramétrica, la distribución de datos subyacente no es conocida con anterioridad. En cambio, algunas técnicas como los histogramas se utilizan para estimar la distribución de datos subyacente y por lo tanto construir el modelo referencia de datos normales que caracteriza el comportamiento de los datos.

2.5.4 Clustering

Los modelos de agrupamiento o *clustering* se utilizan para agrupar patrones similares con características similares. Un clúster es anómalo si es menor o distante de otros grupos en el conjunto de datos (Rassam et al. 2013). Para determinar el patrón de datos pertenecientes a un clúster, se utilizan diferentes medidas de similitud entre ellas la medida de la distancia euclidiana, modelo de geometría de hiper-elipsoides, distribución gausiana y matriz de covarianza.

Esta técnica implica gran carga computacional, para reducir la carga en comunicación se envía solamente un resumen de las medidas de similitud a otros nodos pertenecientes al clúster la reducción toma lugar debido a la agrupación de los datos a nivel local. Además, no se tiene un conocimiento previo de los datos por lo tanto el modelo por *clustering* es completamente no supervisado.

Técnica de *clustering* usando K-means en donde cada nodo local toma un conjunto de datos y obtiene su media y desviación estándar, las envía a un nodo central que compila y procesa las información local y genera un valor global, este valor es enviado al nodo local donde se utiliza para el procesamiento y análisis en la detección de anomalías (Suthaharan et al. 2010).

2.6 Información contextual

La información contextual es considerada como todos los datos disponibles en un ambiente de sensado, no solo los datos procesados por este. La definición de contexto más aceptada es la propuesta por (Dey 2001) “Contexto es cualquier información que puede ser caracterizada para definir una situación de una entidad. Una entidad es una persona, lugar u objeto considerado relevante para la interacción entre un usuario y una aplicación, incluyendo al propio usuario y a la aplicación”.

La situación descrita es el estado del mundo real en un cierto instante de tiempo o durante un intervalo de tiempo en un lugar o espacio específico. En conclusión un contexto es identificado por un nombre, e incluye una descripción de los rasgos característicos de una situación (Schmidt 2002).

Las situaciones pueden ser físicas (localización de una persona) o funcionales (tareas actuales). El contexto se clasifica en cuatro dimensiones: contexto computacional, contexto físico, contexto temporal y contexto de usuario. En un ambiente ubicuo como las redes de sensores (WSN) el contexto físico es el más importante ya que representa todos los parámetros físicos y ambientales, capturados por nodos sensores que son desplegados en una zona de estudio (ejemplos localización del nodo, temperatura, velocidad del viento, nivel de ruido, etc.); Debido a que el contexto físico describe los fenómenos físicos es muy propenso a errores (Konieczny 2012).

Para hacer uso del contexto se debe modelar, siendo este un proceso de representación de la información contextual en la estructura de datos acorde al

conjunto de expresiones o reglas del sistema. Un modelo contextual se establece solo cuando se determinan los objetivos del sistema de conocimiento de contexto. Algunas formas de representación pueden ir de unas simples expresiones relacionando variables, políticas y estructuras XML. La representación del contexto puede incluir clave-valor (una información contextual es representada como una estructura de datos emparejada que incluye un atributo y su valor asociado), el esquema de marcas (representa el contexto en una estructura de datos jerárquica consistente de etiquetas con atributos y contenido), modelos basados en ontologías (representa la información contextual a través de técnicas basadas en semántica), y orientado a objetos (Liu 2014).

Hay que tener en cuenta la existencia de un contexto local referido a un nodo sensor en que sus condiciones y estados de cada subcomponente forman parte de la información contextual local, en el otro lado se encuentra el contexto global que requiere el intercambio de contextos locales entre múltiples nodos, este incluye el sistema (contexto del usuario a nivel de aplicación del sistema) y la red (condiciones de toda la red).

Otra clasificación es por niveles de bajo y alto nivel, el contexto de bajo nivel es inferido de los datos *raw* del sensor, descrito por un elemento o parte de un componente en el nodo, la red o el sistema y el contexto de alto nivel es deducido de múltiples contextos de bajo nivel entre ellos se encuentra el estado del nodo, la red, el sistema (Liu 2014).

El conocimiento de contexto es un prerrequisito para que los sistemas tengan la habilidad de auto-adaptarse a cambios en el comportamiento del usuario o situaciones ambientales.

(Ballari et al. 2009) define cuatro contextos: a) Contexto de organización representa el conocimiento acerca de los objetivos, la seguridad, y las restricciones de privacidad, b) contexto de red genera el conocimiento sobre las funcionalidades, la colaboración y las interrelaciones entre los nodos, c) contexto de nodos ofrece el conocimiento inferido del estado del nodo en un tiempo y el impacto en la interoperabilidad con los otros nodos, d) contexto de sensado, genera el conocimiento sobre el contexto en el que los datos están siendo capturados.

2.7 Brechas existentes

Los trabajos encontrados, han aportado a distintos retos, dejando aún brechas por definir, en el control de calidad de los datos en sistemas de adquisición que hagan frente a la detección de anomalías en datos. Así se hace una clasificación de los trabajos en dos grandes áreas de desarrollo.

1) Mecanismos de control de calidad

En esta categoría se tienen en cuenta dos trabajos muy importante que exploran la detección de anomalías en redes de sensores inalámbricos – WSN.

La evaluación de técnicas realizadas por (Xie et al. 2011) explora primero los aspectos y características más importantes para el diseño de mecanismos de detección de anomalías especialmente para arquitecturas WSN jerárquicas y planas, ya que afirma que las técnicas tradicionales no son aplicables en el contexto de la WSN por la limitación de recursos y que debe existir un balance entre el costo computacional y el costo de comunicación, las técnicas estudiadas se clasifican en: modelos estadísticas, basadas en reglas, minería de datos, inteligencia computacional, teoría del juego, basadas en gráficos e hibridas. (Tabla 2-1).

Categorías	Generalidad	Velocidad	Distribuida	Conocimiento apriori
Estadística	Normal	Normal	Posible	Supuestos
Minería de datos/inteligencia computacional	Alta	Baja	Necesaria	No
Reglas	Baja	Alta	No	Supuestos, experiencia

Tabla 2-1 Evaluación de técnicas de detección de anomalías.

Fuente Anomaly detection in wireless sensor networks: A survey (Xie et al. 2011)

Al observar los resultados de evaluación de las categorías ninguna técnica puede cubrir todos los criterios, donde la técnica de inteligencia computacional es fuerte en el grado de detección y no necesita de un conocimiento previo, pero es débil en la velocidad ya que procesa datos de alta dimensionalidad, en cambio el esquema basado en reglas es rápida en la detección de anomalías, depende de suposiciones o

experiencias anteriores que están sujetas a la intervención humana, pero son muy rápidas ya que maneja poca dimensionalidad de los datos, por último la técnica estadística tiene un balance entre velocidad y el grado de detección, con suposiciones previas, aunque cada una de ellas tiene sus pros y contras, los criterios donde son débiles se pueden solventar agregando o combinando técnicas, atributos o mecanismos.

La revisión realizada por (Rassam et al. 2013) evalúa cada técnica en base a cinco componentes conocido como RODAC (Reducción de dimensionalidad de los datos, detección en línea, detección distribuida, detección adaptativa y exploración de correlación espacial/temporal), los mecanismos evaluados son estadísticos, basados en clasificación, agrupación o *clustering* y vecino más cercano (*nearest-neighbor*).

- Clasificación (Redes Bayesianas, SVM, reglas): Técnicas de clasificación como SVM son costosas computacionalmente, dependientes de parámetros, y, además plantea la necesidad de intervención humana, por consiguiente no aplicables a modelos de detección en línea y flujos de datos altamente dinámicos. En cambio las redes bayesianas exploran las correlaciones y dependencias con atributos de datos, pero no son escalables con datos multivariados.
- Vecino más cercano (*Nearest Neighbor*): el cálculo de la distancia entre los patrones de datos en conjunto de datos multivariados representa gran carga computacional, siendo la escalabilidad de estos modelos una desventaja.
- Modelamiento estadístico: Las técnicas estadísticas son fuertes en la detección de anomalías y eficientes en gestión de recursos (Palpanas 2013). Además si utilizan una ventana de tiempo permite la detección de anomalías en tiempo casi real. Los problemas encontrados es que solo manejan datos univariados, la definición de valores de referencia por ser un ambiente dinámico y cambiante, y la inexistencia de un conocimiento previo de la distribución de los datos(Rassam et al. 2013).
- *Clustering*: Algunos inconvenientes nombrados son: la dependencia para elegir el tamaño del clúster, presenta una alta carga computacional para datos multivariados debido al cálculo de las medidas de distancia entre todos los patrones de datos, las técnicas de agrupamiento no hacen frente a cambios continuos en el flujo de datos, obteniendo un modelo normal de referencia obsoleto al momento de su aplicación.

Nuestra propuesta de evaluación y revisión documental se basó en mecanismo de control de calidad de datos en agricultura de precisión y específicamente los sistemas de adquisición de datos entre ellos las estaciones climatológicas y la WSN; para (Muller et al. 2013; Hubbard et al. 2012; Borghi et al. 2011) las técnicas de control de calidad se realizan en la base de datos sobre un conjunto de datos entregados por las estaciones climatológicas, para (Hamada & Yatagai 2011; Journée & Bertrand 2011; Fiebrich et al. 2010; Estévez et al. 2011) los mecanismo de control de calidad se realizan sobre un solo sensor aplicando técnicas como reglas de decisión y filtros lógicos, explorando la correlación temporal de las instancias de datos.

Los estudios en WSN realizados por (Gwilliams et al. 2012; Elsts et al. 2012; Familiar et al. 2012; Ji et al. 2011; Cid et al. 2011) describen una serie de atributos y parámetros de comunicación utilizados en mecanismos de control de calidad a nivel de red y el enlace de comunicación, muchos de ellos para monitoreo ambiental y definiendo una serie de contextos relacionados con el estado de la WSN. Y por último (Thessler et al. 2011; Christin et al. 2011; Lemmens et al. 2011; Resch et al. 2010) menciona los nuevos desafíos en estándares abiertos para redes de sensores y llama la atención el uso de la información contextual y la iniciativa SWE mejorando el conocimiento que se tiene del ambiente a monitorear, pero no hay una implementación real y además no es utilizado para control de calidad, tan solo ofrece servicios a los usuarios de la aplicación.

2) Información contextual

En (Muller et al. 2013; Hubbard et al. 2012) los metadatos son el resultado de la interacción y el conocimiento de información relacionada con el ambiente y la información procesada de otros sistemas de adquisición llamada información contextual que ayuda a modelar la situación de recolección de datos. Los metadatos provienen de datos del fabricante del sensor, la calibración y mantenimiento del sensor, transferencia de datos confiables a los sistemas de gestión, el despliegue de los sensores en el terreno, localización de la estación, valores de estaciones agroclimatológicos cercanas, patrones de referencia ambientales y agroclimatológicos previstos en la zona y otras características, que por lo general no son tenidos en cuenta en los sistemas de adquisición de datos – SAD.

Para (Konieczny 2012) actualmente la información contextual ya sea características y atributos no es utilizada por los sensores, pero estos pueden ofrecer funcionalidades a la red de sensores, tales como: explorar la integración de información contextual a la pila de protocolos en una WSN (Konieczny 2012); optimizar las tareas de gestión y control de la red (estructuración de la red y ciclos de actividad e inactividad de los sensores) (Liu 2014).

Para (Ngai & Gunningberg 2014; Ji et al. 2011) menciona algunos atributos de información contextual que se evidencian en la etapa de transmisión de datos en un SAD, representados por los parámetros de comunicación, entre ellos: el retardo, el ancho de banda del canal, la capacidad de la red y la velocidad de transmisión.

(Dereszynski & Dietterich 2011) en su técnica de detección de anomalías de datos en un ambiente dinámico plantea un modelo que combina la correlación espacial y temporal incorporando el modelado de un sensor caracterizado por el estado del sensor en cada ventana de tiempo y la observación almacenada de la estación.

(Ballari et al. 2009) estudia la interoperabilidad en los ambientes WSN, para ello hace uso de la información contextual representada como metadatos, definiendo cuatro contextos: a) Contexto de organización, b) contexto de red, c) contexto de nodo, d) contexto de sensado, está la definición mas no la implementación en un ambiente real.

(Liu 2014) hace uso del contexto haga frente al diseño de capas cruzadas en una WSN, el modelado de la información contextual se realiza por medio de una ontología, se diseña la arquitectura del nodo totalmente definida y simulada en la plataforma OPNET.

Las brechas existentes que se identificaron son:

- Para los sistemas de adquisición ya sea WSN o estaciones climatológicas el procesamiento de los datos para mejorar su confiabilidad están orientados a un sensor específico, abarcando solamente anomalías presentes en la red que hacen uso de parámetros de comunicación; los mecanismos propuestos de detección de anomalías utilizan una arquitectura centralizada realizados en su gran mayoría en la base de datos o centro de datos.
- Las propuestas del uso de información contextual en los sistemas de adquisición son insipientes y no hay evidencias de su uso para mejorar la calidad de los datos;

Se tienen identificados elementos de contexto que son utilizados en otros dominios de aplicación y específicamente en la WSN para proveer servicios de localización y calidad en el ancho de banda del enlace de transmisión, pero hasta ahora se está estudiando cómo implementarlo adecuadamente.

- Las propuestas arquitectónicas en la integración entre el modelamiento de información contextual y el proceso de control de calidad de los datos in-situ son escazas y exploratorias, dejando la posibilidad de proponer marcos de referencia para este tipo de integración.

2.8 Resumen

En este capítulo una introducción a conceptos relacionados con agricultura de precisión, detección de anomalías y sistemas de control de calidad de datos han sido presentados. Específicamente como se aborda desde los sistemas de agricultura de precisión la etapa de adquisición de datos y ya que estos son susceptibles a fallas, errores y anomalías, además, se presentan una serie de estudios sobre detección de anomalías en WSN con diferentes técnicas y finalizando con un análisis de brechas en los dominios de mecanismos de control de calidad de datos en un sistema de adquisición de datos y de información contextual inmersa en el ambiente de agricultura de precisión. Las brechas de investigación que han sido discutidas se abordarán como contribución de este trabajo en el Capítulo 3 y 4.

Capítulo 3

Mecanismos de procesamiento de información

La propuesta presentada en este trabajo, se refiere a los sistemas de adquisición de datos presentes en una aplicación de agricultura de precisión y dentro de ella se aborda la etapa de adquisición y procesamiento de los datos, teniendo en cuenta las diferentes fallas y anomalías presentes en estos sistemas por ser limitados en recursos, el estudio se centra en un control de calidad de los datos que involucra la detección de anomalías a nivel de datos: parámetros fuera de rango, valores atípicos, fallas constantes y ruido excesivo.

A continuación se presenta el escenario de control de calidad de datos para agricultura de precisión, los mecanismos propuestos y la evaluación de los mismos para validar el grado de detección de anomalías.

3.1 Control de calidad de los datos para agricultura de precisión

En la agricultura de precisión los parámetros ambientales y físicos que son objeto de estudio son variados entre ellos están: temperatura, humedad, velocidad y dirección del viento, presión atmosférica, radiación solar, etc. Con características y atributos propios que definen el fenómeno con el mayor detalle posible evidenciando un tipo de datos multivariados.

El monitoreo en una área de estudio implica el conocimiento previo de las condiciones donde son desplegados los sensores, más exactamente la localización y condiciones atmosféricas reinantes en la zona, definiendo una correlación espacial que implica

modelar un contexto que representa una situación o una condición que provee un conocimiento de los atributos y propiedades ambientales en el área de estudio.

La captura de información la realiza el sistema de adquisición, en donde los sensores son la fuente de datos de un flujo continuo, llamado datos RAW, estas instancias de datos se recopilan en intervalos de tiempo periódicos, definidos al momento de configuración de la red y dependientes del parámetro a ser monitoreado. La dependencia de una instancia de datos con el tiempo de muestreo supone una correlación temporal entre el valor de la muestra en el tiempo t (x_t) y el valor de la muestra en el tiempo t+1 (x_{t+1}).

Los datos RAW son procesados y almacenados en pequeños buffer temporales, para su posterior envío a un centro de datos, el tiempo entre la captura de los datos y su transmisión es muy pequeño, casi en tiempo real y depende en sí mismo de la limitación de los recursos en el sistema de adquisición.

La decisión de cuando enviar los datos, depende del sistema de adquisición estructurando la red en una topología jerárquica, estableciendo un tiempo de transferencia, hacia el centro de datos donde se reciben y son almacenados de manera permanente en una base de datos.

Los datos enviados al centro de datos no sufren algún tipo de procesamiento, siguen siendo los datos RAW capturados, no se envía información adicional relacionada con el proceso de captura de los datos, el procesamiento efectuado, la calidad de las muestras y otros atributos relevantes que serán tenidos en cuenta en una tarea de análisis de información.

Los datos agroclimatológicos son usados por sistemas de agricultura de precisión que toman decisiones relacionadas a la aplicación de fertilizantes, el control de plagas y enfermedades, las necesidades de agua en el cultivo, etc. Que al no recibir gran cantidad de datos confiables son asumidos como valores ambientales verdaderos afectando la productividad y calidad de las cosechas, y por ende las condiciones económicas de los pequeños agricultores. Un ejemplo de aplicación de agricultura de precisión se observa en la Figura 3-1.

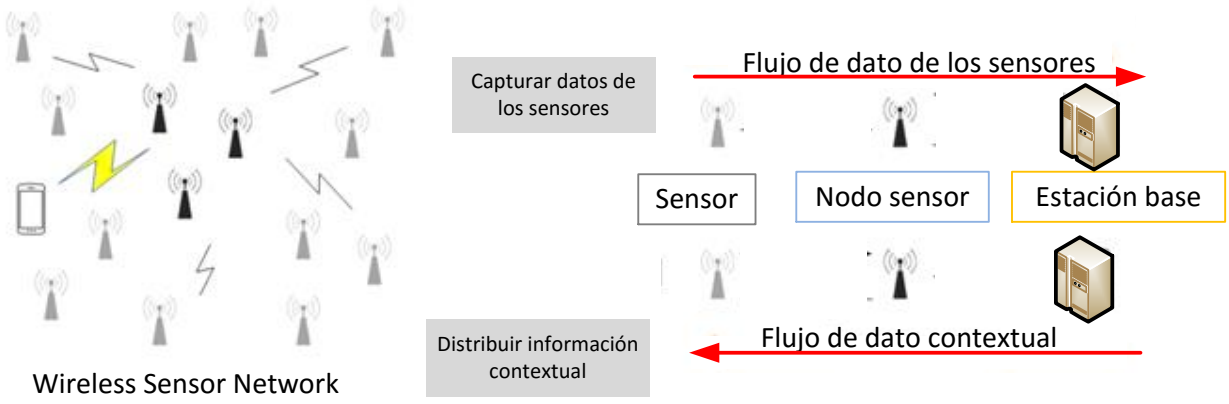


Figura 3-1 Modelo de comunicación.

Como se vio anteriormente, estos sistemas de adquisición de datos son susceptibles a presentar anomalías. Esta propuesta se enfoca en anomalías de datos y la estrategia para detección de anomalías y obtener datos confiables es realizar mecanismos de control de calidad de datos dentro de los nodos a nivel local, para evitar el envío de datos erróneos al centro de datos.

Para cada tipo de anomalía se realiza un mecanismo de control de calidad de los datos, es de aclarar que dichos mecanismos deben ser los más livianos posibles teniendo una relación directa entre la capacidad de procesamiento y el consumo de energía ya que se ejecutan en el sitio de adquisición de los datos y a medida que lleguen las muestras (tiempo real).

Para esta propuesta los mecanismos de control de calidad de datos seleccionados en una aplicación de agricultura de precisión están basados en técnicas estadísticas, conjugado con reglas de decisión, aunque estos mecanismos presentan ventajas y desventajas, existen una serie de consideraciones en su aplicación y la forma como son abordadas las deficientes de estas técnicas, se tiene lo siguiente:

- ✓ Datos multivariados, evaluar las series de tiempo de las muestras de un mismo tipo, recopiladas en una ventana de tiempo muy pequeño como una primera etapa del proceso y un análisis entre varios tipos de variables de diferente tipo en una etapa posterior, categorizando el proceso de control de calidad de datos en niveles.
- ✓ La existencia de una correlación temporal entre instancias de datos, ya que se capturan flujos continuos de datos RAW que son procesados en tiempo real.

- ✓ Definición de valores de referencia por ser un ambiente dinámico y cambiante: es aquí donde se tiene el mayor aporte y es el uso de información contextual a nivel local y de bajo nivel que permita que el mecanismo pueda ser adaptable a dichos cambios, a partir de la serie de datos o los datos RAW que están siendo capturados.
- ✓ Inexistencia de un conocimiento previo de la distribución de los datos, es importante tener técnicas incrementales y adaptativas que a medida que los datos son analizados permitan establecer el comportamiento de las señales y por ende ayudan a modelar la distribución de los datos. Para esta propuesta se supone que las señales analizadas tienen un comportamiento de una distribución normal aproximada.
- ✓ La dependencia de suposiciones o experiencias anteriores para establecer valores de umbral, supone la intervención humana, pero la inserción de información contextual representada por medio de metadatos, siendo el resultado del análisis y procesamiento de datos históricos presentes en la zona de estudio y en épocas muy similares. Ejemplo insertar valores de temperatura media, máxima y mínima de un lugar cercano donde van a ser desplegados los sensores para el tiempo de duración del estudio, esto ayuda a establecer límites o umbrales que mejoran la calidad de los datos.
- ✓ Los datos controlados, que son datos generados del proceso de control de calidad, están etiquetados, para conocer con mayor detalle el reporte de la anomalía y si este dato puede ser usado para mejorar el proceso de control de calidad.
- ✓ Actualización de parámetros: es el propio sistema que infiere la actualización, por medio de un aprendizaje por reforzamiento que establece la realimentación de los datos procesados y controlados por calidad, hacia la información contextual definida con anterioridad, representada por medio de metadatos.
- ✓ Existe un problema la mayoría de técnicas de detección de anomalías no distinguen entre errores, eventos y ataques maliciosos, los mecanismos usados en esta propuesta tampoco lo hacen dejando ese estudio para trabajos futuros. Logramos un pequeño acercamiento con el uso de la información contextual donde se busca correlacionar dos variables usando datos que han sido controlados previamente.

En conclusión el mecanismo propuesto reduce la dimensionalidad de los datos, detecta en línea o en tiempo real anomalías, se adapta a las condiciones cambiantes del flujo de datos capturado, y explora la correlación temporal y espacial; todo esto con la ayuda de un proceso de enriquecimiento obtenido de la información contextual,

representada por metadatos y producida dentro del mismo nodo o insertada de una fuente externa como resultado de un análisis en un centro de datos. (Tabla 3-1).

Técnicas	Dimensionalidad de datos		Estructura del modelo			Modo de operación		Adaptación a cambios		Correlación	
	Univariados	Multivariados	Local	Centralizada	Distribuida	En línea	Fuera de línea	Adaptativa	No adaptativa	Espacial	Temporal
Clustering		X			X		X		X		
Vecino más cercano		X			X		X		X		
Clasificación		X			X	X		X		X	X
Estadística	X				X		X		X	X	X
Propuesta		X	X			X		X		X	X

Tabla 3-1 Resumen de características de la propuesta.

Se pueden utilizar métodos estadísticos como paso previo para determinar el método más apropiado para un aprendizaje supervisado (Meratnia & Havinga 2010). Ya que limpia datos univariados de flujos continuos de datos por medio de umbrales predefinidos, estos valores de referencia se actualizan por medio de distribuciones de probabilidad locales y en otras ocasiones se reciben desde una estación central.

Correlación espacial y temporal

- ✓ Para manejar las correlaciones se traslada el problema a la detección de anomalías puntuales y se desacoplan las correlaciones con la información contextual.
- ✓ La correlación temporal en donde se evidencia la diferencia en las muestras de datos del mismo sensor en diferente instante de tiempo. Se chequea la consistencia temporal de los valores medidos con base a unos umbrales, resultado del análisis climatológico propuesto por (Vejen et al. 2002) (National Oceanic and Atmospheric Administration 1998). Para datos RAW y un intervalo de muestreo de 10 segundos la variabilidad límite permitida para la temperatura es de 2°C, la humedad relativa es de 5% y la presión atmosférica de 0.3 hPa. Para datos procesados se define una variabilidad máxima permitida para valores de temperatura mayores de 3°C se marca como sospechoso y para 4°C como erróneo, la humedad relativa a partir del 10% como sospechoso y el 15% como erróneo y una variabilidad mínima requerida de los datos instantáneos supone un valor mínimo requerido para la temperatura de 0.1°C y la humedad relativa de 1%.
- ✓ La correlación espacial vista por la diferencia entre las muestras de datos de diferentes sensores en el mismo tiempo. Las anomalías espaciales para temperatura y humedad casi no se presentan ya que estos parámetros poseen baja variación espacial (Fiebrich & Crawford 2001). Pero se va a utilizar la información contextual generada por medio de un análisis de las condiciones meteorológicas de la zona de estudio, en los meses anteriores o históricos, dicha información establece unos límites superiores e inferiores que reflejan un rango admisible para evaluar los valores a nivel local y es introducido al proceso por un agente externo en su configuración inicial y/o cada cierto periodo de tiempo, para nuestra propuesta este periodo es mensual.

Etiquetado de datos de salida

El etiquetado de datos según la clasificación de la OMM se realiza con marcas de calidad, el dato resultante del proceso de control de calidad se llama dato controlado y se establece una etiqueta llamada marca de calidad, la cual refleja el estado de confiablidad del dato, y busca identificar la causa de la falla, entre ellas tenemos:

Q.1. Buenos (datos con errores menores o iguales a un valor especificado) datos que han pasado todas las pruebas.

Q.2. Inconsistentes (uno o más parámetros son incompatibles) identifica valores atípicos que están dentro del rango de valores esperados, pero se desvían de otras instancias de los datos. Además evalúa el ruido excesivo que presentan las muestras en la serie temporal.

Q.3. Dudoso (sospechoso) son valores que están dentro de los límites físicos del sensor pero no corresponde con valores ambientales de la zona de estudio, evaluando una correlación espacial.

Q.4. Erróneos (datos con errores superiores a un valor especificado) esta marca identifica anomalías de datos que se desvían significativamente de un rango de valores esperados.

Q.5. Perdidos o faltantes, indica la falta de una instancia de datos que no se capturo o el proceso de captura fallo, debido a una falla en el sensor evidenciando un valor con poca o nula variabilidad para un periodo de tiempo más largo de lo esperado, puede estar dentro o fuera del rango esperado.

3.2 Mecanismos de control de calidad de datos seleccionados

Los mecanismos definidos para el control de calidad de los datos se establecen en dos partes: la primera de ella hace uso de reglas de decisión que utiliza parámetros o umbrales definidos, desde la información del fabricante del sensor y parámetros ambientales de la zona, representados como metadatos que proviene de la información contextual de una fuente externa; la segunda parte con la ayuda de técnicas estadísticas livianas se detectan anomalías de datos que se encuentran dentro de unos valores aceptables, pero sus desviaciones en la serie de tiempo no son normales, actualizando parámetros para adaptarse a las condiciones actuales de los valores capturados y etiquetando las instancias de datos como normales o anómalas.

1) Reglas de decisión contextual

Estas reglas utilizan información contextual del fabricante del sensor, como el límite máximo permitido (δ_{max}) y el límite mínimo permitido (δ_{min}) de la temperatura del sensor, si se encuentra fuera de este rango de valores, se etiqueta la muestra como errónea y se lleva un conteo de las instancias de datos erróneas. (Algoritmo 1).

Algoritmo 1 Regla Contextual Información del Fabricante

Entrada: Dato Raw x_i
Salida: Dato Controlado x_c, Marca de calidad mk_{qa}

1: **si** $x_i \geq \delta_{min}$ y $x_i \leq \delta_{max}$ **entonces**
2: $x_c \leftarrow Bueno$
3: **si no**
4: $x_c \leftarrow Falla$
5: $Count_{erróneo} + 1$
6: $mk_{qa} \leftarrow MKQ\_ERRÓNEO$ {x_i Etiquetado como Erróneo}
7: **fin si**

Para la siguiente regla se toma la información ambiental de la zona, representada por un máximo medio (λ_{max}) y un mínimo medio (λ_{min}) proveniente de datos históricos. Los valores de datos que están por fuera de este rango ambiental se etiquetan como dudosos y se incrementar un contador de instancias de datos dudosas. (Algoritmo 2).

Algoritmo 2 Regla Contextual Información Climatológica

Entrada: Dato Raw x_i
Salida: Dato Controlado x_c, Marca de calidad mk_{qa}

1: **si** $x_i \geq \lambda_{min}$ y $x_i \leq \lambda_{max}$ **entonces**
2: $x_c \leftarrow Bueno$
3: **si no**
4: $x_c \leftarrow Falla$
5: $Count_{dudoso} + 1$
6: $mk_{qa} \leftarrow MKQ\_DUDOSO$ {x_i Etiquetado como Dudoso}
7: **fin si**

Después de aplicar dichas reglas se generan nuevos datos procesados como la media (μ) y la desviación estándar (σ) de las muestras, para adaptarse a cambios de la señal o a la distribución de los datos en el espacio temporal, utilizamos el algoritmo de desviación estándar incremental (Algoritmo 3). En el cálculo de la media (μ) no se utilizan los valores etiquetados como erróneos; ya que son datos que no pasan la prueba de rango admisible y el dato se debe eliminar del conjunto de datos.

Algoritmo 3 Desviación Estándar Incremental

Entrada: Valor de la muestra x

Salida: σ desviación estándar

1: Inicializar la media μ con el valor de la muestra x

$$\mu \leftarrow x$$

2: Inicializar la desviación estándar σ

$$\sigma \leftarrow 0$$

3: Calcular la media μ, donde i es el indice de la muestra actual x_i

$$\mu_i = \mu_{i-1} + \frac{x_i - x_{i-1}}{i} \quad (1)$$

2) Funciones umbral

Una regla sencilla para descartar valores dudosos consiste en excluir de la serie de datos (N muestras), el valor que se encuentra fuera de un rango especificado. Con los datos restantes se calculan nuevamente la media (μ) y la desviación estándar (σ). El rechazo puede considerarse justificado si la desviación del valor sospechoso (x_s) con respecto a la media (μ) es, por lo menos, cuatro veces la desviación estándar (σ) de los valores (Hubbard et al. 2012). (Ecuación 3.1).

$$|x_s - \mu| \geq= 4\sqrt{\frac{1}{N-1}\sum_{i \neq s}(x_i - \mu)^2} = 4\sigma' \quad \text{Ecuación 3.1}$$

3) Técnicas estadísticas

Estas técnicas hace uso de una ventana de tiempo o un número fijo de muestras, sobre la cual se generan unos valores observados que son comparados con valores definidos en una tabla de datos con diferentes grados de confianza de 90%, 95% y 99%. Entre las técnicas utilizadas esta la prueba T y la prueba Q, definidas para detectar valores atípicos.

- ✓ Prueba Tn (distribución continua y tasa de muestreo constante) el test de paso o variabilidad máxima permitida. La prueba T_n se define en la Ecuación 3.2

$$T_n = \frac{|x_s - \mu|}{\sigma}$$ Ecuación 3.2

En este caso, vale destacar que la media (μ) y la desviación estándar (σ) son calculados incluyendo el valor dudoso. El rechazo está justificado si el valor T_n calculado es mayor a los que se indican en la Tabla 3-2.

N	90%	95%	99%
3	1,15	1,15	1,15
4	1,46	1,48	1,49
5	1,67	1,71	1,75
6	1,82	1,89	1,94
7	1,94	2,02	2,10
8	2,03	2,13	2,22
9	2,11	2,21	2,52
10	2,18	2,29	2,41

Tabla 3-2 Valores de T_n para distintos porcentajes de confianza.

✓ Prueba Q

En la prueba Q se determina una Q_{exp} para el dato sospechoso que luego sería comparado con los valores de Q dados en la Tabla 3-3. El Q_{exp} se define en Ecuación 3.3:

$$Q_{exp} = \frac{|x_s - x_n|}{W}$$ Ecuación 3.3

Donde x_n es el valor más próximo al valor sospechoso (x_s) y W es la dispersión. Si Q_{exp} es mayor al Q dado en la Tabla 3-3, el dato dudoso puede ser rechazado con el grado de confianza indicado. A continuación se presenta una tabla para valores de Q a distintos porcentajes de confianza.

N	90%	95%	99%
3	0,941	0,970	0,994
4	0,765	0,829	0,926
5	0,642	0,710	0,821
6	0,560	0,625	0,740
7	0,507	0,568	0,680
8	0,468	0,526	0,634
9	0,437	0,493	0,598
10	0,412	0,466	0,568

Tabla 3-3 Valores de Q para distintos porcentajes de confianza.

4) Prueba pendiente y tendencia

Hace uso de reglas de decisión sobre la serie de tiempo, en donde el valor actual de la muestra de datos es comparada con la anterior y si la diferencia supera un límite establecido (λ_{max}) con anterioridad, el valor actual de la muestra es identificada como sospechosa, si el valor está por debajo de un límite inferior (λ_{min}) el dato es marcado como perdido. Los límites establecidos se basan en información contextual proporcionada por datos históricos referentes a variables climatológicas propuesta por la OMM, sugiriendo que la variabilidad entre instancias de datos no supere un máximo y un mínimo establecido a nivel ambiental, detectando un ruido excesivo en los datos. (Algoritmo 4).

Algoritmo 4 Regla de Decisión de Gradientes Climatológicos

Entrada: Dato Procesado x_p
Salida: Dato Controlado x_c, Marca de calidad mk_{qa}

si $|x_p - x_{p-1}| \geq \lambda_{max}$ **entonces**
  $x_c \leftarrow Falla$
  $Count_{dudoso} + 1$
  $mk_{qa} \leftarrow MKQ\_DUDOSO$ {x_p Etiquetado como Dudoso}
si no, si $|x_p - x_{p-1}| \leq \lambda_{min}$ **entonces**
  $x_c \leftarrow Falla$
  $Count_{perdido} + 1$
  $mk_{qa} \leftarrow MKQ\_PERDIDO$ {x_p Etiquetado como Perdido}
si no
  $x_c \leftarrow Bueno$
fin si

Cálculo de fallas contantes, lo que se busca es determinar pocas o nulas variaciones para un periodo de tiempo más largo de lo esperado en una serie de tiempo, en esta técnica se establece un contador de muestras ($Count_{repet}$) que se incrementa al detectar un valor constante ($|x_p - x_{p-1}| == 0$) y si el contador llega a un límite (γ) este es marcado como perdido. (Algoritmo 5).

Algoritmo 5 Prueba de Falla Constante

Entrada: Valor de la muestra x_p
Salida: Dato Controlado x_c, Marca de calidad mk_{qa}

```
Inicializar el contador de valores constantes Count_repet ← 0
si |x_p − x_{p−1}| == 0 entonces
    Count_repet + 1
    si Count_repet ≥ γ entonces
        x_c ← Falla
        Count_repet ← 0
        Count_perdido + 1
        mk_qa ← MKQ_PERDIDO {x_p Etiquetado como Perdido}
    si no
        x_c ← Bueno
    fin si
si no
    x_c ← Bueno
fin si
```

5) Prueba de consistencia Interna

Se relacionan dos parámetros, en este caso relacionando las instancias de datos de los sensores con el nivel de la batería, si esta prueba falla se marcan los datos como anómalos.

El nivel de la batería se verifica datos admisibles comparando la información contextual proporcionada por el fabricante y la medida del nivel de la batería en ese instante, marcando los datos como buenos o anómalos. (Tabla 3-4).

Nivel de Batería	Datos Controlados	Resultado
Anómalo	Anómalo	Anómalo
Anómalo	Bueno	Anómalo
Bueno	Anómalo	Anómalo
Bueno	Bueno	Bueno

Tabla 3-4 Mecanismo de consistencia interna.

3.3 Evaluación de los mecanismos de control de calidad de datos

Para evaluar los mecanismos propuestos usaremos métricas de clasificación para determinar entre errores correctos versus errores incorrectos, entre ellas sensibilidad y especificidad, al ser evaluado un conjunto de datos en varios escenarios que presentan diferentes tipos de fallas modelado en el software de simulación de redes OMNET. Sumado a esto, se etiquetan las instancias de datos como buenos, inconsistentes, dudosos, perdidos y erróneos. Y por último se realiza el análisis correspondiente de los resultados.

3.3.1 Evaluación propuesta

Métricas de evaluación de algoritmos

Las métricas de evaluación propuestas hacen uso de ciertos términos entre ellos: los verdaderos positivos (TP), los verdaderos negativos (TN), los falsos positivos (FP) y los falsos negativos (FN). (Tabla 3-5).

TP (*True Positive*)	Valores anómalos clasificados por el sistema como anómalos.
TN (*True Negative*)	Valores buenos que no han sido detectados como anomalías.
FP (*False Positive*)	Valores buenos clasificados como anomalías.
FN (*False Negative*)	Valores anómalos pero no detectados por el sistema

Tabla 3-5 Relación entre la condición esperada y el resultado de la prueba.

La métrica utiliza dos parámetros de rendimiento para cada escenario o experimento: la sensibilidad (Se) y la especificidad (Sp).

- ✓ La sensibilidad indica el porcentaje de anomalías que han sido detectadas correctamente sobre el total de anomalías.

$$S_e = \frac{T_p}{T_p + F_n}$$

- ✓ La especificidad indica el porcentaje de datos buenos que han sido detectados correctamente sobre el total de datos buenos ingresados.

$$S_p = \frac{T_n}{T_n + F_p}$$

- ✓ Precisión en la detección, proporciona una idea de cómo muchos de los valores que se etiquetan como anómalos son verdaderamente fallas del sensor.

$$P = \frac{T_p}{T_p + F_p}$$

- ✓ La relación de falsas alarmas – FAR, es el número de datos buenos clasificados como anómalos (falsos positivos), dividido por el número total de datos anómalos reportados.

$$F = \frac{F_p}{T_n + F_p}$$

Un buen esquema de detección posee alta precisión en la detección, con baja tasa de falsas alarmas, además de minimizar la utilización de recursos que determina mayor velocidad de detección pero probablemente pérdida en efectividad (Xie et al. 2011). Siendo la efectividad de la detección representada por la precisión en la detección, tasa de detección (sensibilidad) y falsas alarmas.

Etiquetado de datos

El etiquetado de datos se realiza marcando los datos al final de la ejecución de cada mecanismo de control de calidad, como resultado del mismo proceso. Las marcas definidas son:

Q.1. Buenos (MKQ_BUENO)
Q.2. Inconsistentes (MKQ_INCONSISTENTE)
Q.3. Dudoso (MKQ_DUDOSO)
Q.4. Erróneos (MKQ_ERRONEO)
Q.5. Perdidos o faltantes (MKQ_PERDIDO)

Escenarios de fallas

La propuesta es insertar fallas de datos (valores atípicos, ruido excesivo, fallas contantes y valores fuera del rango) en el conjunto de datos para que el mecanismo clasifique los datos como buenos o anómalos, y por medio de las métricas de evaluación analizar su resultados, entre los escenarios (Farruggia 2011) a evaluar tenemos 1) Fallas continuas se produce durante todo el experimento; por ejemplo, un sensor produce una salida constante, o las lecturas del sensor son alteradas por un ruido gaussiano. 2) Fallas discontinuas ocurren a intervalos regulares de tiempo; asumiendo que en estos intervalos el sensor defectuoso genera una salida constante, mientras que retorna de un funcionamiento normal. Estas fallas discontinuas se caracterizan por dos parámetros: la duración de la falla, y el número total de apariciones durante el experimento. En resumen la propuesta de escenarios de falla es el siguiente:

Sin Fallas

1. Ideal sin ningún tipo de falla.

Fallas constantes (*Struck-at*)

2. Falla Continua con valor constante de 19°C.
3. Falla Discontinua con valor constante igual a la muestra anterior.

Fallas con ruido gaussiano (*Outliers* y *Spike*)

4. Falla Continua con ruido gaussiano representado con la función Normal (4, 3.0) °C.
5. Falla Discontinua con ruido gaussiano representado con la función Normal (4, 2) °C.

Fallas con valores fuera de rango (*Out-of-range*)

6. Falla Discontinua con valores fuera de rango entre 30 °C + función Normal (4,3) °C.

3.3.2 Ambiente de prueba

Se diseña un escenario de simulación de un sistema de adquisición de datos agroclimatológicos, en la plataforma de simulación de redes OMNET, junto a un conjunto de datos (*dataset*) de temperatura proporcionado por Intel Berkeley Research

Lab (IBRL) (IBRL 2004) que ha sido capturado por una red de sensores inalámbricos WSN en los laboratorios de investigación de Intel en la universidad de Berkeley. La red consiste de 54 nodos sensores Mica2Dot, desplegados durante 30 días entre el 28/02/2004 hasta 05/04/2004. Cuatro tipos de mediciones han sido capturadas: luz, temperatura, humedad y voltaje de la batería. El tiempo de muestreo es de 31 segundos. En esta propuesta se utiliza el parámetro de temperatura, el nivel de voltaje de la batería y los tiempos de muestreo, sobre el nodo sensor identificado como N1.

3.3.3 Modelamiento del nodo sensor y control de calidad

La implementación del modelo (Figura 3-2), se conforma en grupos funcionales, donde algunos de ellos generan paquetes o mensajes de información, que son enviados a otros bloques donde se realiza la detección de anomalías y se etiquetan los datos, por último se reportan los datos como buenos o anómalos, además incluye un módulo que induce fallas a los datos generados.

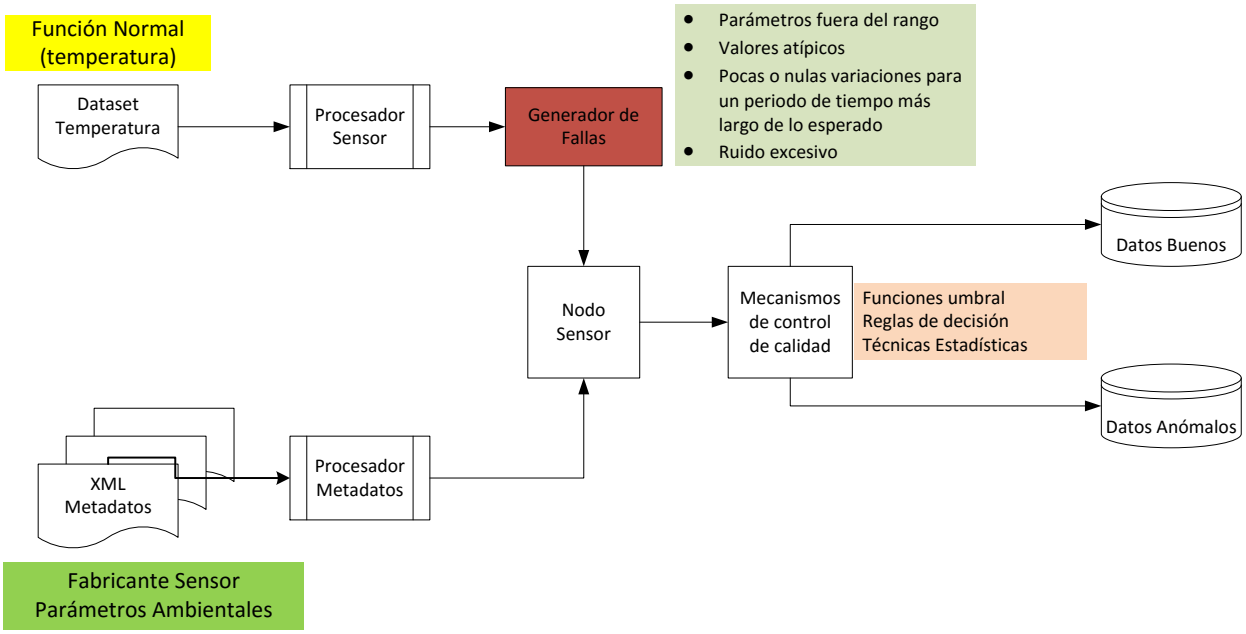


Figura 3-2 Modelado ambiente de adquisición e inserción de fallas en OMNET.

Procesador del sensor: su función principal es generar los paquetes de información del sensor, simulando la capturada de los datos por medio de recorrer el conjunto de datos disponible en un archivo de texto.

Procesador de metadatos: establece los umbrales o valores de referencia del dato del fabricante y de datos climatológicos de la zona, que se envía periódicamente al nodo sensor.

Generador de fallas: hace uso de funciones estadísticas e intervalos de tiempo para inducir fallas en el conjunto de datos y enviar paquetes de datos contaminados al nodo sensor.

Nodo sensor: se implementa el proceso de control de calidad de los datos para valores fuera de rango, recibiendo las instancias de datos, aplicando las funciones de umbrales con la ayuda de los metadatos de fabricante y ambiental, entregando datos procesados y controlados a las próximas etapas.

Mecanismos de control de calidad: la aplicación de técnicas estadísticas a los datos controlados, con el fin de detectar anomalías de falla constante, ruido excesivo y valores atípicos, generando nueva información contextual, etiquetando los datos para el reporte de anomalías y clasificando las muestras como buenas o anómalas.

Datos buenos y anómalos: almacén temporal de los datos como resultado del proceso de control de calidad.

3.3.4 Modelamiento de metadatos

Se realiza un proceso de abstracción de la información más relevante del fabricante del sensor de temperatura, para definir valores aceptables del sensor, además se extrae información climatológica de la zona de estudio comprendida en el periodo de estudio, dicho proceso se menciona a continuación:

- ✓ Modelamiento de los metadatos del fabricante del sensor, para ello en los datos suministrados por los propietarios del conjunto de datos se especifica el tipo de sensor utilizado en las mediciones de temperatura y se extrae los datos más importante, a partir de allí realizamos una implementación en XML para ser consumidas por el modelo de metadatos (Tabla 3-6).

Parameters	Conditions	Minimum	Typical	Maximum
Humidity				
Resolution		1%RH	1%RH	1%RH
			8 Bit	
Repeatability			±1%RH	
Accuracy	25℃		±4%RH	
	0-50℃			±5%RH
Interchangeability	Fully Interchangeable			
Measurement Range	0℃	30%RH		90%RH
	25℃	20%RH		90%RH
	50℃	20%RH		80%RH
Response Time (Seconds)	1/e(63%)25℃, 1m/s Air	6 S	10 S	15 S
Hysteresis			±1%RH	
Long-Term Stability	Typical		±1%RH/year	
Temperature				
Resolution		1℃	1℃	1℃
		8 Bit	8 Bit	8 Bit
Repeatability			±1℃	
Accuracy		±1℃		±2℃
Measurement Range		0℃		50℃
Response Time (Seconds)	1/e(63%)	6 S		30 S

```
<?xml version="1.0" encoding="UTF-8"?>
<sensor>
   <profile id="DHT11">
      <maxHum>80</maxHum>
      <minHum>20</minHum>
      <accurHum>0.05</accurHum>
      <maxTemp>50</maxTemp>
      <minTemp>0</minTemp>
      <accurTemp>2</accurTemp>
      <samplesRate>1</samplesRate>
   </profile>
   <profile id="DHT22">
      <maxHum>100</maxHum>
      <minHum>0</minHum>
      <accurHum>0.02</accurHum>
      <maxTemp>125</maxTemp>
      <minTemp>-40</minTemp>
      <accurTemp>0.5</accurTemp>
      <samplesRate>2</samplesRate>
   </profile>
</sensor>
```

Tabla 3-6 Modelamiento metadatos del fabricante del sensor.

- ✓ Modelamiento de los metadatos ambientales de la zona de estudio (Tabla 3-7); en este caso, con la ayuda de portales climatológicos (U.S. Climate Data n.d.), se extrae la información de promedios históricos de temperatura de la ciudad de Berkeley Estados Unidos en el mes de marzo del año 2004, ya que el conjunto de datos se capturo entre el 28 de febrero y el 5 de abril de 2004. Esta información es almacenada en un archivo XML, que puede ser modificado por el proceso de control de calidad.

Temperatura máxima media	71.7°F (normal: 64°F) – 22.05°C (17.78°C)
Temperatura mínima media	48.4°F (normal: 46°F) – 9.11°C (7.78°C)
Temperatura media	60.05°F (normal: 55°F) – 15.58°C (12.78°C)

Tabla 3-7 Metadatos de temperatura histórica de la zona de Berkeley, Estados Unidos.

3.3.5 Modelamiento de fallas

En el modelamiento de las fallas de ruido y valores atípicos, se tiene en cuenta la información climatológica ofrecida por la OMM, en donde se establece que la temperatura media horaria suele tener una distribución normal en climas tropicales y una distribución algo más asimétrica en latitudes medias, las temperaturas medias

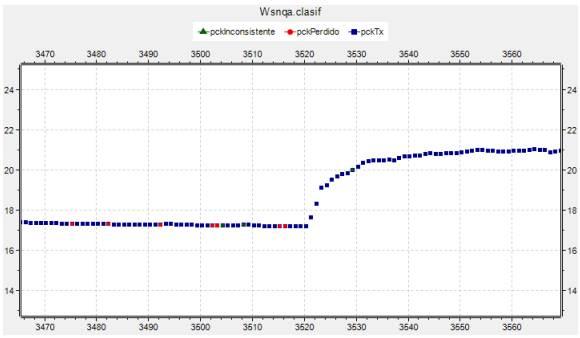
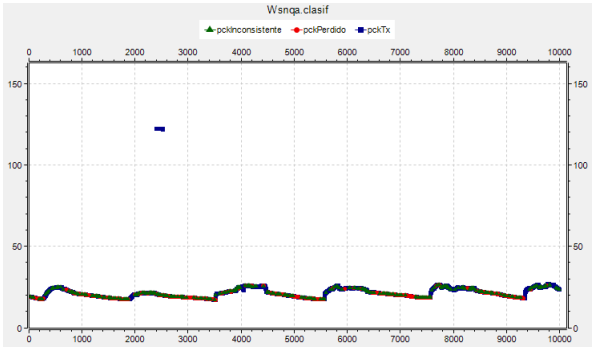
diarias muestran una distribución casi normal, en cambio las máximas diarias presentan una distribución asimétrica positiva principalmente en verano. Por el contrario las temperaturas mínimas diarias presentan una distribución asimétrica negativa sobre todo en invierno. Teniendo en cuenta lo anterior el modelamiento del ruido se basa en una distribución Normal (μ, σ) en donde (μ) es la media y (σ) la desviación estándar. Ejemplo Normal (4, 3.0) °C y Normal (4, 2) °C.

Las fallas discontinuas se basan en insertar durante un tiempo establecido (duración de la falla) un ruido de tipo gaussiano y a intervalos de tiempo periódicos. Para los escenarios propuestos la duración de la falla es de 5 periodos de muestreo de los datos de los sensores y el intervalo que ocurre este proceso es cada 20 periodos de muestreo de la señal. Para nuestro caso tenemos un conjunto de datos de 10000 muestras se inyectan alrededor de 2500 datos anómalos.

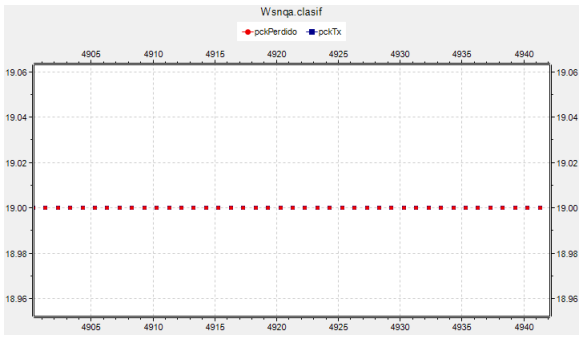
El modelamiento de fallas con valores fuera de rango se establece un valor constante de 30 °C y se sobrepone una distribución Normal (4,3) °C.

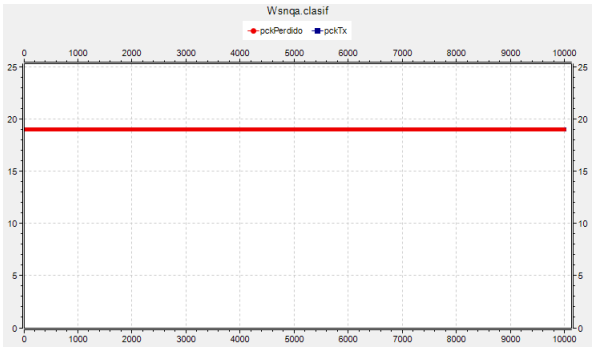
3.3.6 Resultados

A continuación se presenta el análisis de los mecanismos de control de calidad ante los diferentes escenarios de fallas propuestos. En la Figura 3-3 se muestra a la derecha las gráficas del conjunto de datos seleccionado con 10000 muestras para cada uno de los escenarios de prueba y a la izquierda una imagen ampliada del tipo de falla modelada y etiquetada por medio de las marcas de calidad presentadas anteriormente y evaluadas por el mecanismo de control de calidad propuesto.

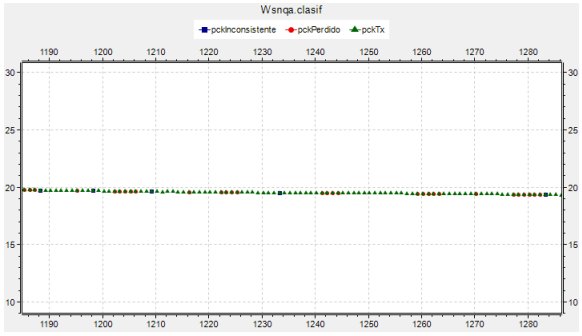
Ideal sin ningún tipo de falla	
	

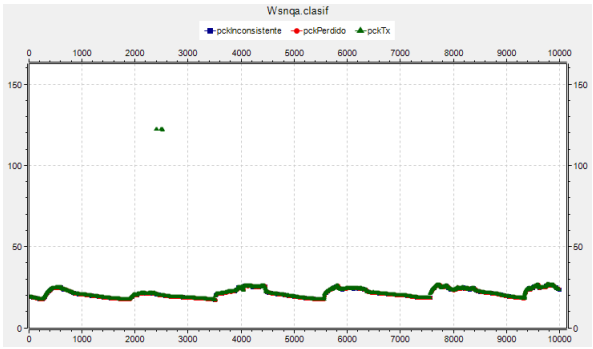
Falla Continua con valor constante de 19°C.



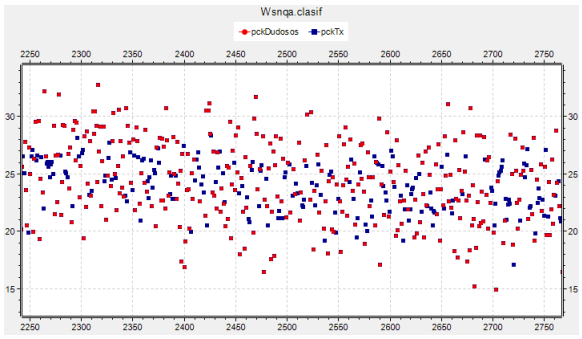


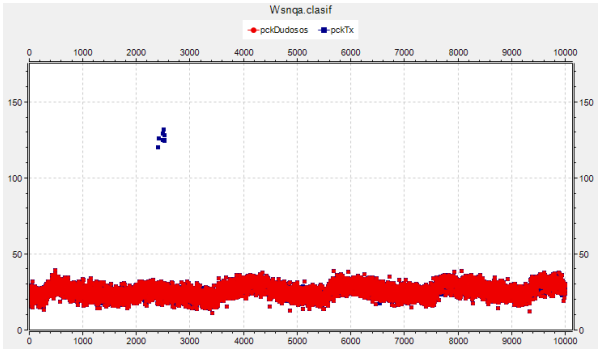
Falla Discontinua con valor constante



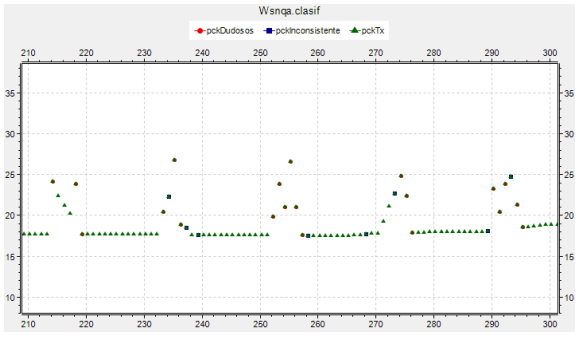


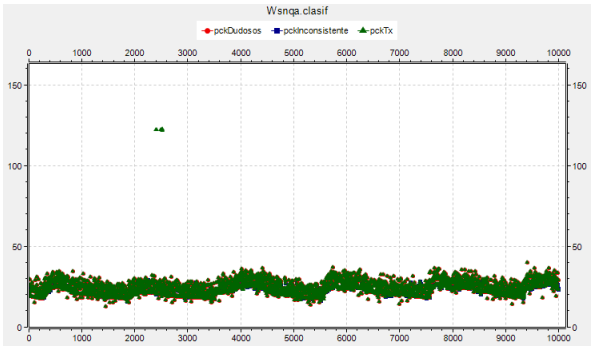
Falla Continua con ruido gaussiano con función Normal (4,3) °C





Falla Discontinua con ruido gaussiano con función Normal (4,3) °C





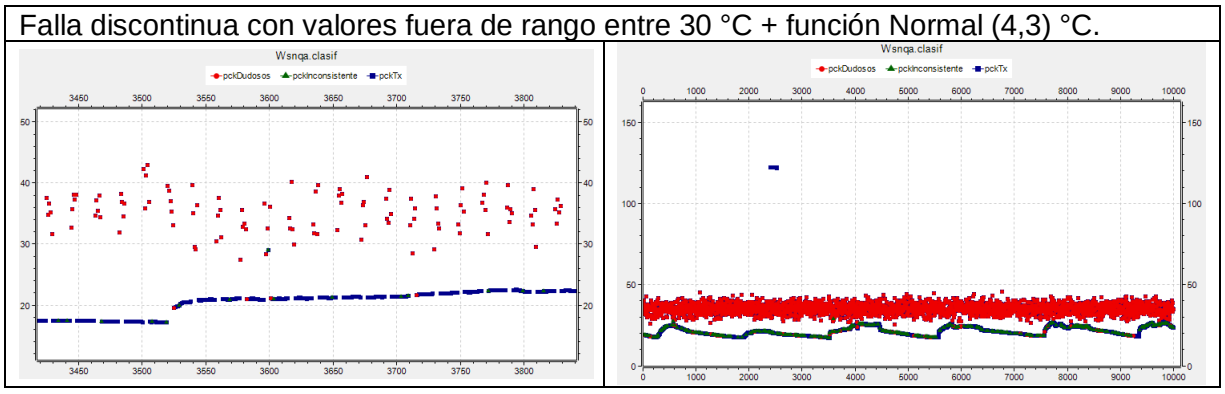


Figura 3-3 Escenarios de fallas y marcas de calidad en el conjunto de datos.

En la Tabla 3-8 se muestra los resultados del proceso de simulación del mecanismo de control de calidad para cada uno de los escenarios propuestos, aquí se evidencia que en un escenario de falla con un valor constante ya sea continuo o a intervalos de tiempo definidos con anterioridad (discontinuo). La etiqueta MKQ_PERDIDO representa el mayor número de datos anómalos, esto significa que el mecanismo de calidad tiene implementado un algoritmo que verifica y controla dichas fallas constantes.

En cuanto a fallas de ruido y valores atípicos la etiqueta MKQ_DUDOSO, presenta el mayor número de datos anómalos, siendo el mecanismo de control de calidad un rango contextual basado en parámetros climatológicos y complementado con una técnica estadística que filtra los valores que están dentro del rango anteriormente establecido etiquetando los datos anómalos como MKQ_INCONSISTENTE. Es de aclarar que la aplicación de ciertos mecanismos sobre el conjunto de datos de prueba, también realiza su proceso de control de calidad, generando falsos positivos, razón por la cual el número de datos anómalos es superior a la cantidad de valores con falla insertados (2500 datos con falla de ruido).

El algoritmo de técnica estadística es muy robusto para detectar anomalías, es por ello que se utiliza en una capa interior del proceso, para que sea más efectivo y no se contamine con valores fuera de rango o fallas constantes. Por ultimo en las fallas con valores fuera del rango, se detectan datos anómalos por medio del mecanismo de funciones umbral, en este caso los parámetros o rangos climatológicos de la zona de estudio son los principales en la detección de anomalías, además se etiquetan como MKQ_DUDOSO.

Como resumen del análisis, los diferentes mecanismos de control de calidad ayudan en la detección de un tipo de falla o anomalías, es por ello que se deben combinar para obtener mejores resultados en la detección, además el etiquetado del proceso ayuda a visualizar e identificar el tipo de falla presentada y así tomar las respectivas acciones a ser corregidas en un proceso posterior.

Otro aspecto a tener en cuenta es la selección de los valores de los parámetros, más que todo los contextuales ya que están expuestos a cambios constantes según la zona de estudio y las situaciones ambientales en un intervalo de tiempo, por eso se necesita que estos parámetros presenten una característica de adaptación según el tipo de señal que están recibiendo de sus sensores.

Escenario	TX	MKQ BUENO	MKQ INCONSISTENTE	MKQ DUDOSO	MKQ ERRONEO	MKQ PERDIDO
Ideal	9991	8818	342	1	8	822
Continua con valor constante	9991	1	0	1	0	9989
Discontinua con valor constante (2500)	9991	6798	356	3	12	2822
Continua con ruido gaussiano Normal(4,3)	9991	2946	668	6368	8	1
Discontinua con ruido gaussiano Normal(4,2)	9991	6846	670	1960	8	507
Discontinua con ruido gaussiano Normal(4,3)	9991	6592	731	2153	8	507
Discontinua con valores fuera de rango	9991	6411	373	2682	7	512

Tabla 3-8 Resultado etiquetados de los escenarios de falla propuestos.

En la Tabla 3-9 se puede observar que la detección de anomalías constantes en cuanto a sensibilidad es bastante alta ya que el mecanismos de control de calidad utilizado se basa en discriminar valores que no presentan mucha variabilidad entre muestras y se reportan la mayoría de fallas introducidas. Por otra parte la detección de valores fuera de rango presenta una sensibilidad alta, ya que utiliza técnicas basadas en reglas contextuales que son rápidas y de alto grado de detección. Las fallas discontinuas con ruido gaussiano afectan la sensibilidad cuando el ruido introducido es muy similar a los valores de los datos capturados por los sensores, muestra de ello es el 78.9% de sensibilidad cuando hay ruido definido con una función normal de media 4 y desviación estándar 2, comparado con el 89.1% de ruido gaussiano con media 4 y desviación estándar 3. Por último la de menor sensibilidad es un ruido constante mezclado en los datos de los sensores con un 58.8%.

En cuanto a especificidad en todos los escenario se evidencia un 84.4%, demostrando que nuestro mecanismo de control de calidad detecta la ausencia de anomalías en datos buenos o normales.

En cuanto a precisión, el mecanismo propuesto tiene mayor precisión cuando las anomalías son continuas, ya sea con fallas constantes (100%) y con ruido gaussiano (83.4%). Para los escenarios con fallas discontinuas la precisión baja hasta el 62% en promedio, esto se debe al uso de la ventana de tiempo en el mecanismo de control y la forma como se planteó el escenario de fallas ya que al ingresar las fallas discontinuas se genera una serie temporal de valores que entre ellos son normales o buenos, pero comparada con valores buenos y en otra ventana de tiempo da como resultado anomalías.

La relación de falsas alarmas indica que en promedio un 16% de los datos buenos son etiquetados como anómalos, sobretodo en el escenario discontinuo. Para el escenario continuo no se presenta dicha relación ya que consideramos que toda la muestra está contaminada con valores anómalos.

Escenario	Sensibilidad	Especificidad	Precisión	Falsas Alarmas
Ideal				
Continua con valor constante	100,0%		100,0%	
Discontinua con valor constante (2500)	80,0%	84,1%	62,6%	15,9%
Continua con ruido gaussiano Normal(4,3)	58,8%		83,4%	
Discontinua con ruido gaussiano Normal(4,2)	78,9%	84,4%	62,7%	15,6%
Discontinua con ruido gaussiano Normal(4,3)	89,1%	84,4%	65,5%	15,6%
Discontinua con valores fuera de rango	94,9%	83,9%	66,4%	16,1%

Tabla 3-9 Resultados de evaluación: sensibilidad, especificidad, precisión y falsas alarmas.

En conclusión el mecanismo de control de calidad propuesto es efectivo ya que presenta una aceptable sensibilidad con un promedio del 80% y una especificidad con el 84%.

Si evaluamos la precisión con un promedio de 62% y la sensibilidad con un promedio de 80%, concluimos que nuestro mecanismo de control de calidad clasifica de manera aceptable datos anómalos y datos buenos.

3.4 Resumen

Los mecanismos de control de calidad de datos agroclimatológicos propuestos es una combinación de reglas de decisión que son rápidas, y permiten detección en tiempo real, sumadas con técnicas estadísticas que son confiables y presentan una gran tasa de detección, la utilización de ambas técnicas se modelan en un ambiente de simulación llamada OMNET, que nos permitió identificar y definir un algoritmo de control de calidad de datos, al usar un conjunto de datos no etiquetados, al mismo tiempo seleccionar ciertos parámetros de referencia, la identificación de información contextual de fuentes externas y el etiquetado de datos de salida con marcas de calidad.

Al final se tiene un conjunto de mecanismos evaluados y validados con métricas de especificidad por encima del 84%, la precisión del 62%, sensibilidad del 80% y tasa de falsas alarmas del 15.6%. Además, se resalta la importancia de utilizar la información contextual externa y se evidencia la necesidad de actualizar los parámetros de referencia para mejorar el mecanismo de control de calidad de datos.

Capítulo 4

Arquitectura de referencia

En este capítulo se describe la arquitectura del sistema propuesto, teniendo en cuenta la detección de anomalías, la información contextual de fuentes externas y la información contextual generada en el propio sistema para actualización de parámetros de referencia y que el proceso de control de calidad sea un ciclo de mejora continuo. Al inicio se definirán algunos modelos que están implícitos en un sistema de adquisición y otros que permitan cumplir con las funcionalidades de un control de calidad, enseguida se muestran los componentes de la capa de control de calidad, su modelo lógico y por ultimo sus diferentes niveles de abstracción en donde están inmersos los mecanismos de control de calidad definidos anteriormente.

4.1 Modelos y arquitectura de alto nivel

La arquitectura propuesta a través de sus componentes persigue generar en el SAD una mejora en la calidad de los datos capturados de los sensores y enviarlos al centro de datos como información de alta confiabilidad para la toma de decisiones a nivel agrícola. En el proceso de definir una arquitectura se establecen las siguientes funcionalidades:

- Una mejora en la calidad de los datos recolectados en los sensores.
- Acceso automático a los parámetros y capacidades de los sensores.
- Recuperación en tiempo real de los datos y el despliegue de servicios en la web.
- Interoperabilidad de estándares de la OGC y la OMM.

Dicha arquitectura se concibe desde cuatro aspectos: el primero de ellos el modelo físico de un sistema de adquisición representado en un nodo sensor, el segundo un modelo contextual (Ballari et al. 2009) que enriquece el proceso de control de calidad de los datos, el tercero un modelo interoperable, orientado a servicios y conectado a la web como es SWE (*Sensor Web Enablement*) especificado por la OGC (*Open Geospatial Consortium*) (Percivall et al. 2007), y el cuarto un modelo de aprendizaje que permita adaptar ciertas acciones del proceso de control de calidad a partir de la información contextual a nivel local y actualizar ciertos parámetros de referencia que se utilizan en los mecanismos de control de calidad.

4.1.1 Físico

El modelo físico se basa en la estructura hardware de un nodo sensor que se compone de la unidad de sensado, la unidad de procesamiento, la unidad de comunicación y la unidad de energía. La unidad de sensado recopila los datos de los sensores y los transfiere a la unidad de procesamiento, esta a su vez posee un pequeño almacenamiento que guarda los datos temporalmente durante las tareas de procesamiento. La unidad de comunicación transfiere los datos desde el nodo a otros nodos o estación base y por último la unidad de energía que suministra a los sensores y al nodo la fuente de energía para realizar sus funciones. Se debe tener en cuenta las restricciones de estos componentes hardware para mantener un grado de autonomía aceptable.

Según (Oliveira & Rodrigues 2011) existen diferentes funcionalidades que debe ejecutar un nodo sensor entre ellas: adquisición de datos y acondicionamiento de señales para diferentes sensores; ii) almacenamiento temporal de los datos capturados; iii) procesamiento de los datos; iv) análisis de los datos procesados para diagnóstico y generación de alarmas; v) auto monitoreo en cuanto a la fuente de alimentación; vi) programación y ejecución de tareas de medición; vii) gestión de la configuración del nodo sensor; viii) recepción, transmisión y reenvió de paquetes; y ix) gestión y coordinación de las comunicaciones y la red.

4.1.2 Contextual

El contexto global se refiere a un contexto que no es deducido localmente, sino que proviene de fuentes externas, tal como los centros de datos donde se generan datos históricos de la zona, parámetros ambientales mínimos, medios y máximos, etc.

El contexto local (Ballari et al. 2009) describe la información contextual deducida localmente: un contexto de sensado que ofrece el conocimiento relacionado con la captura, evaluación y comprensión de los datos capturados de los sensores; un contexto de nodo que evalúa el estado del nodo en un tiempo específico y el impacto de interoperabilidad con otros nodos; el contexto de red referida a las funcionalidades de colaboración e interrelación con otros nodos y por último el contexto a nivel de la organización que impone los objetivos, la seguridad y las restricciones de privacidad de la organización; estos contextos buscan enriquecer el proceso de control de calidad de los datos definiendo un conjunto de metadatos. (Figura 4-1).

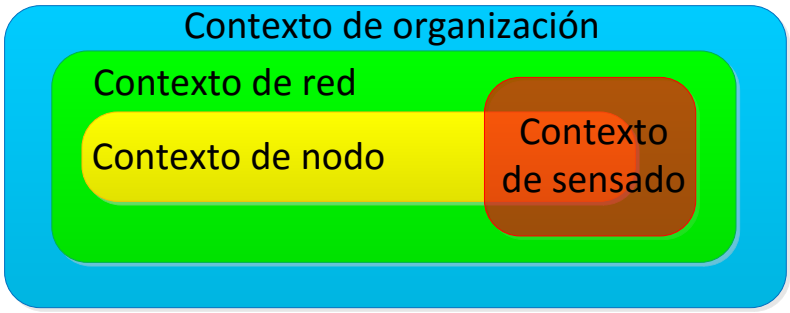


Figura 4-1 contextos para interoperabilidad en WSN.

Fuente: Metadata behind the interoperability of wireless sensor networks (Ballari et al. 2009)

4.1.3 Interoperable

El modelo interoperable se basa en *Sensor Web Enablement* - SWE (Lemmens et al. 2011) que aporta un *framework* de estándares y buenas practicas a nivel del modelo de información con acceso a los parámetros y capacidades del sensor desde un entorno web; SWE abarca: el esquema de observaciones y medidas (*Observation & Measurement Schema*, O&M) (Cox 2011) que define un modelo de datos y esquemas XML para representar observaciones y medidas obtenidas por un sensor. El lenguaje de modelamiento de sensor (*Sensor Model Language, SensorML*) (Botts & Robin 2007) que define un modelo de datos y esquemas XML para describir sistemas y procesos de sensores; y el servicio SOS (Na & Priest 2007) que define una interfaz de servicio web para la petición, filtrado y lectura de información de observaciones y sistemas de sensores. Los estándares SWE permiten adicionar metadatos para describir las características técnicas del sensor, precisión, condiciones, escenarios de medición y observaciones.

4.1.4 Aprendizaje por reforzamiento

Un sistema de control de calidad debe permitir un ciclo de mejora continua es allí donde se utiliza un enfoque de aprendizaje continuo y el más apropiado es el aprendizaje por refuerzo ya que no construye modelos explícitos o preferencias en selección de acciones que pueden ser optimas o medio optimas, además usa métricas de desempeño para modelar un ambiente complejo y se basa en el conocimiento del contexto. Según (Yau et al. 2012) el agente de aprendizaje se comunica con su propio entorno pero no con un entrenador, para cada acción del agente, el medio ambiente responde con una recompensa, que representa la eficacia de la acción en ese instante de tiempo; Sin embargo, no existen acciones "correctas" o "incorrectas". El objetivo es encontrar una política, que seleccione una acción en cualquier instante de tiempo, que conduzca a la mejor recompensa posible del medio ambiente. (Figura 4-2).

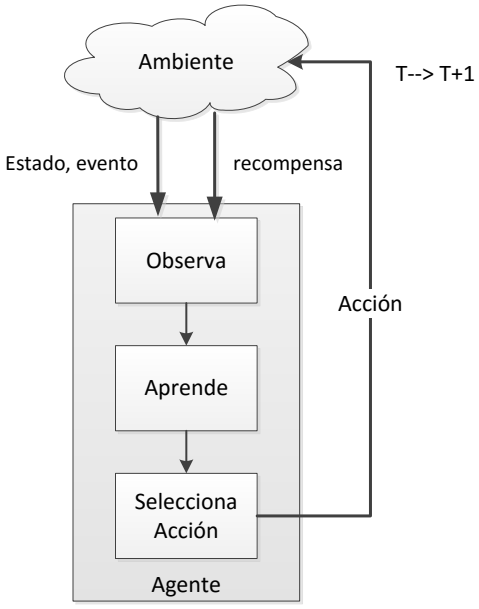


Figura 4-2 Modelo de agente de aprendizaje por refuerzo en el ambiente.

En la Figura 4-3 se muestra la arquitectura de alto nivel propuesta, en donde el modelo contextual se sobrepone sobre el modelo físico con el fin de evidenciar los metadatos que son descritos para ser interoperables según la iniciativa SWE.

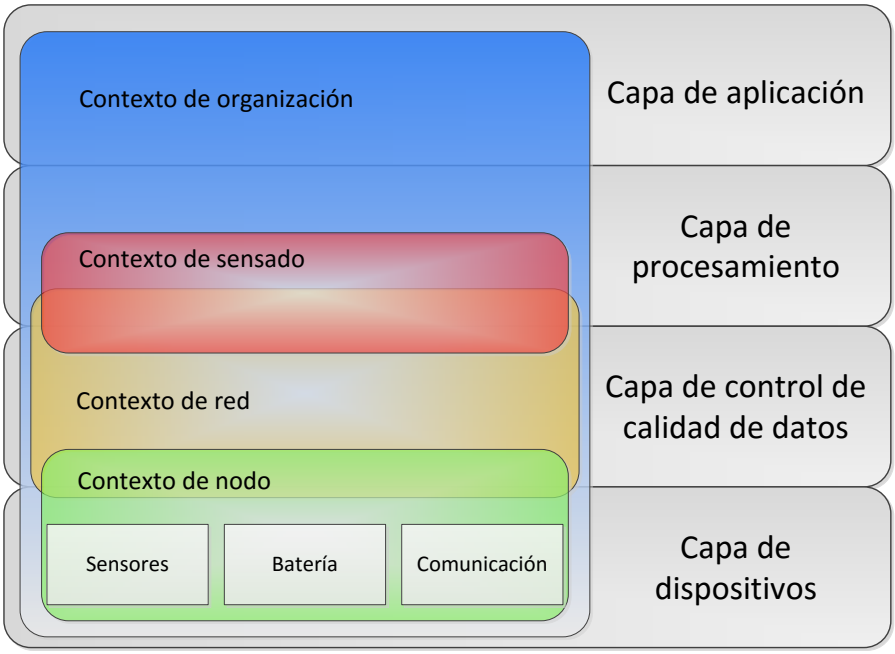


Figura 4-3 Arquitectura Contextualizada de Alto Nivel de un SAD.

A continuación se describe cada componente de la arquitectura propuesta.

- El contexto de nodo y la capa de dispositivos se refiere a la interacción con los dispositivos físicos o hardware del sistema en ellos se encuentra los diferentes sensores agroclimatológicos (temperatura, humedad, radiación solar, precipitación, velocidad y dirección del viento), ya sean de naturaleza analógica o digital, la tecnología de comunicaciones a emplear (802.15.4, Zigbee, Bluetooth, WiFi, GSM/GPRS) y la fuente de energía (batería y paneles solares). Dentro de los metadatos que podemos encontrar se refiere a atributos de la fuente de energía (estado de la energía), atributos de la capacidad de la memoria (nivel de almacenamiento), atributos del modo de operación y protocolos del módulo de comunicación (servicios ofrecidos, calidad del enlace, topología de la red). Al final provee el conocimiento de las condiciones del sistema en un tiempo determinado teniendo en cuenta el estado de los dispositivos conectados y sus componentes, para ofrecer datos de alta calidad.
- La capa de control de calidad de los datos asegura que los datos capturados reflejen fielmente las condiciones reinantes en la zona y que exista una coherencia entre los elementos observados en ese instante de tiempo. En esta capa deben existir una serie de mecanismos que ayuden a reducir el impacto que tiene los errores y fallas de los datos capturados en la calidad de los datos y es aquí que

confluyen el contexto de sensado, nodo y red ofreciendo mayor conocimiento local y global que ayuden a evaluar y entender las condiciones del sistema cuando los datos son capturados por los sensores.

- El contexto de sensado unido a la capa de procesamiento involucra todas las operaciones y condiciones de captura de datos teniendo en cuenta la información temporal (tiempo de muestreo) y el tipo del fenómeno a ser observado (temperatura, humedad, etc.), esto lo realiza el sistema operativo (tareas actuales en ejecución) que además, gestiona el uso de la memoria (ventanas de tiempo para almacenamiento temporal de los datos) y da formato a los datos para ser enviados a un centro de datos.
- El contexto de organización y la capa de aplicación buscan ofrecer servicios e interoperabilidad a otros usuarios siguiendo las políticas de la organización en cuanto a seguridad, privacidad y entrega de información confiable, haciendo uso de estándares abiertos como el SWE.

4.2 Modelo lógico de control de calidad de los datos

El Control de Calidad de Datos (CCD) se refiere a los procesos y técnicas enfocados a mejorar la eficacia de los datos existentes. Además, debe incluir procedimientos para el retorno a la fuente de datos para verificarlos y prevenir la repetición de los errores. Es decir, si el valor de temperatura ambiente para la ciudad de Popayán Colombia registrado en el SAD es de 42°C y según promedios históricos de temperatura para la ciudad es de 19°C, este valor esta por fuera del patrón de referencia ambiental, el dato debe ser marcado como inconsistente y se llevaran a cabo las acciones necesarias, como puede ser la calibración del sensor, el mantenimiento del sensor debido a un posible ruido en la medición de la variable temperatura producto de un agente externo.

El conocimiento de los procedimientos relativos al proceso de captura de los datos y al control de calidad permite a los usuarios evaluar la validez de la observación y convertir el proceso en un ciclo de mejora continua. Es por ello que se construye un modelo lógico (Figura 4-4) que describe el flujo de los datos, las actividades para mejorar la calidad y los resultados obtenidos del proceso.

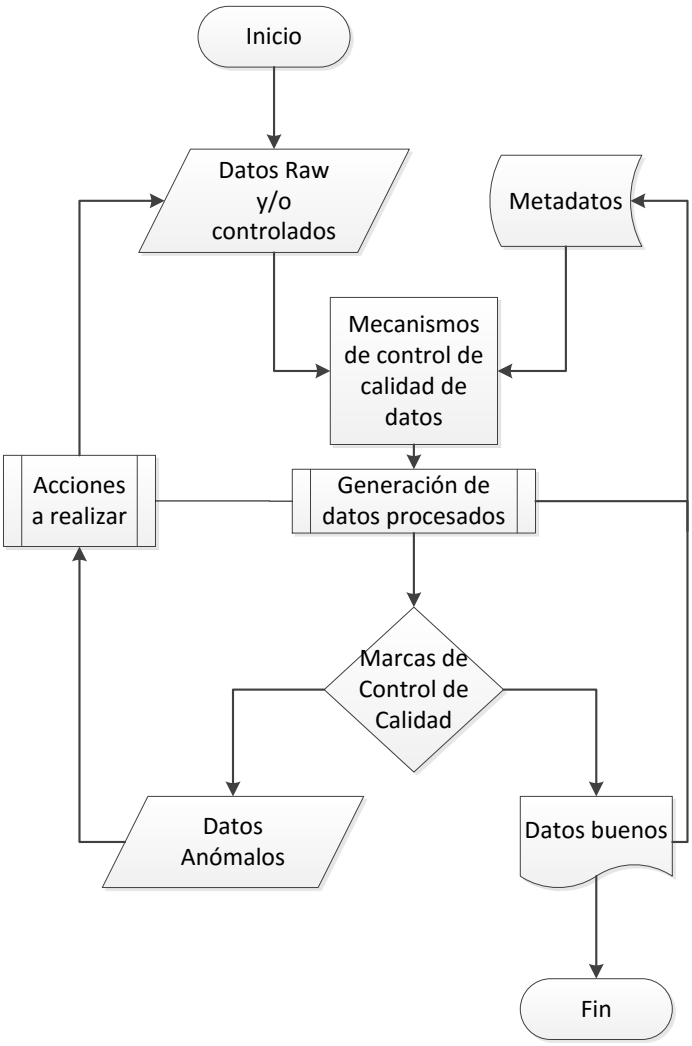


Figura 4-4 Modelo lógico capa de control de calidad de datos.

Un modelo de datos que está dividido en dos partes como lo sugiere (Mason et al. 2014), la primera parte los datos brutos y datos controlados (Huang et al. 2014) que son recolectados de los sensores o vienen de etapas anteriores y la otra parte un conjunto de metadatos estándar o atributos de información entre ellos: datos del fabricante del sensor, métodos, esquemas de adquisición de datos, etc. Además se generan nuevos conjuntos de metadatos que son necesarios para informar a los futuros usuarios del tipo de datos que genera el sistema, la manera en que se capturaron los diversos conjuntos de datos y todos los problemas inherentes a la adquisición. Todos modelados bajo los estándares descritos por SWE, siguiendo las normas de la OMM.

Los mecanismos de control de calidad para las diferentes fases del proceso de recolección de datos, son los responsables de mantener la confiabilidad de los datos, reducir en lo posible los errores, verificar y validar los datos capturados con el conjunto de metadatos. Entre los mecanismos se encuentran: reglas de decisión contextual, funciones umbral que se basan en parámetros de referencia climatológica, filtros lógicos, técnicas estadísticas como la prueba T y la prueba Q, entre otros.

Las marcas de control de calidad como lo sugiere la WMO (Organizacion Meteorológica Mundial - OMM 2011) ofrecen una categorización de los datos en: a) buenos, b) inconsistentes, c) dudoso o sospechoso, d) erróneos y e) perdidos o faltantes.

La generación de datos procesados de todo el conjunto de datos y la ventana de tiempo establecida ayuda a concebir nuevos datos llamados información contextual local, entre ellos está la media, la desviación estándar, los valores máximos y mínimos de la señal y los valores máximos y mínimos de diferencias entre muestras consecutivas en la serie temporal. Los datos que ayudan a generar esta información contextual provienen de datos etiquetados como buenos, dudosos y perdidos.

Los datos buenos son el resultado de un dato que ha superado satisfactoriamente el proceso de control de calidad y los datos que no superan el control de calidad son considerados como anómalos (erróneos, dudosos, perdidos e inconsistentes), los datos erróneos no participan en futuros cálculos. La salida del sistema de control de calidad son datos controlados que debe incluir indicadores sobre el resultado de la medición si es correcta o incorrecta, así como un conjunto de afirmaciones del resumen sobre los sensores, las marcas de calidad que se han atribuido a la observación; el historial de las modificaciones introducidas en los valores y en cualquier marca de calidad conexa. Además se debe conservar los valores de los datos originales, estimados y modificados.

Las acciones a realizar son llevadas a cabo si los datos son marcados como anómalos (erróneos, sospechosos, perdidos), en ella los procedimientos evalúan los datos en una escala temporal y espacial para verificar los valores permisibles, la homogeneidad meteorológica y la verosimilitud física, aumentando la calidad del sistema y reduciendo la incertidumbre que depende del nivel donde se encuentre la calidad de los datos.

Entre las acciones a tomar esta la de remplazar o conservar los datos anómalos, algunas alternativas son:

- Alternativa I, Ignorar: un dato anómalo no se sustituye, ni se modifica.
- Alternativa II, Medir nuevamente: reemplazar un dato anómalo realizando una nueva medida al sensor y cuyo dato supere el proceso de control de calidad.
- Alternativa III, Reemplazar por la media: sustituir un dato anómalo por el valor calculado de la media a partir de los datos anómalos y normales.
- Alternativa IV, Interpolar: sustituir un dato anómalo con el valor obtenido mediante la interpolación de los datos inmediatamente anterior y siguiente en la serie temporal.

4.3 Componentes capa de control de calidad

Los modelos de arquitectura propuestos, sumado al modelo lógico de control de calidad, imponen una serie de requisitos al sistema que deben implementarse para lograr un mejor desempeño en la mejora de datos con las limitaciones de recursos que presenta un nodo sensor y enriquecidos con información contextual global y local. A continuación se enumeran algunos requisitos:

- ✓ Recuperar información contextual local (dentro del proceso) y global (fuentes externas) y representarla como metadatos.
- ✓ Ejecutar mecanismos de control de calidad.
- ✓ Almacenamiento de datos y parámetros de referencia al interior del sistema de control de calidad.
- ✓ Generar salida de datos controlados en calidad y etiquetados con marcas de calidad, buscando la interoperabilidad para ser transferidos a un centro de datos o estación base.
- ✓ Mecanismo de actualización o aprendizaje de parámetros de referencia, que ayuden a definir acciones sobre los datos anómalos y mejoren el proceso de control de calidad.

Los requisitos demandan el diseño de ciertos componentes (Figura 4-5) que estructuran la capa de control de calidad.

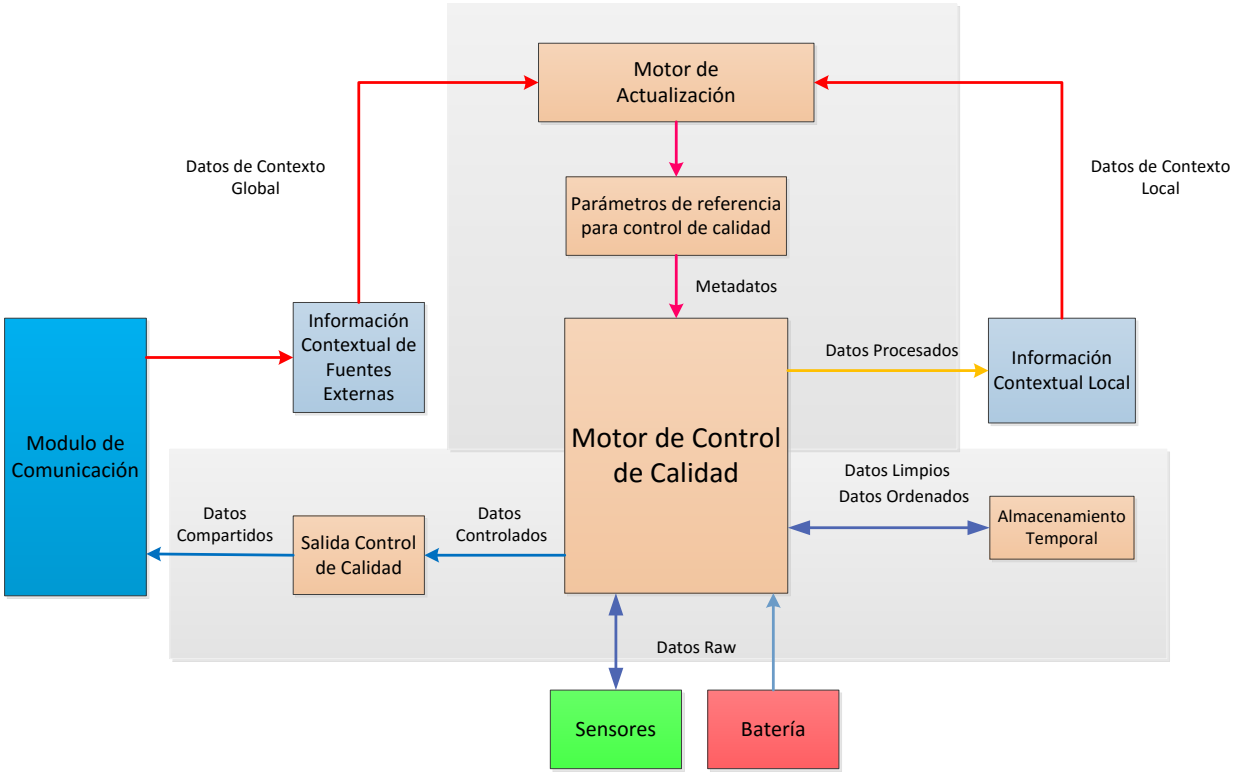


Figura 4-5 Nodo sensor con mecanismo de control de calidad.

Módulo de comunicación: es la interfaz radio que nos permite comunicar el sistema de adquisición - SAD con el centro de datos o estación base, este módulo tiene dos vías una vía que recibe información de fuentes externas para distribuir el contexto global y la otra vía envía los datos capturados por los sensores (datos *raw*) que han pasado por el proceso de control de calidad (datos controlados); que se desean compartir (datos compartidos) con otros sistemas.

Información contextual de fuentes externas: es la información generada por datos climatológicos que han sido procesados por otros medios, más que todo en un centro de datos y se convierten en datos históricos y parámetros de referencia ambiental. Además de información relacionada con la zona objeto de estudio, el fabricante del sensor y las características de los componentes hardware que integran el SAD.

Parámetros de referencia: es la información validada y adaptada que proviene del contexto local y global para ser consumida por el motor de control de calidad, representada como metadatos y estructurada siguiendo el estándar SensorML de la SWE.

Motor de control de calidad: ejecuta los mecanismos de control de calidad con base en los parámetros de referencia (metadatos), los datos capturados por los sensores (datos *raw*) y los algoritmos de procesamiento de control de calidad, este decide la lógica de control de calidad de los datos y genera datos controlados que se envían a la estación base o centro de datos y además genera datos procesados (información contextual local) que enriquecen el proceso de control de calidad.

Información contextual local: se compone de datos procesados que es generada por los mecanismos de control de calidad, convirtiéndose en un punto de referencia para actualizar los parámetros en su proceso de adaptación a cambios externos.

Motor de actualización: con base en la información contextual local y global produce un afinamiento de parámetros de referencia, que permite generar un ciclo de mejora en el proceso de control de calidad, aquí está diseñado el aprendizaje por reforzamiento, mencionado en los modelos anteriores.

Sensores: es el componente que captura valores de fenómenos climatológicos como la temperatura, humedad, precipitación, etc. para ser consumidos por los mecanismos de control de calidad, en sí, es la fuente original de datos climatológicos sin ningún tipo de procesamiento (datos *raw*).

Batería: fuente de alimentación del sistema de procesamiento, sistema de comunicación y los diferentes sensores, el valor medido (datos *raw*) con un nivel de bajo voltaje repercute en datos inconsistentes y erróneos.

Almacenamiento temporal: permite generar una ventana de tiempo para que el motor de control de calidad pueda ejecutar ciertos mecanismos que requieren datos ordenados y datos limpios para mejorar la calidad.

Salida control de calidad: es el resultado de los datos de los sensores expuestos al proceso de control de calidad, generando etiquetas con marcas de calidad llamados datos controlados. Dichos datos son compartidos y enviados por el módulo de comunicación a la estación base, utilizando estándares interoperables como el O&M y SOS definidos en SWE.

4.4 Niveles de control de calidad

A continuación se muestra en la Figura 4-6 una jerarquía que se establece en la capa de control de calidad de los datos en un SAD producto del contexto de los datos y del sistema. Todos los niveles sigue el mismo modelo lógico propuesto anteriormente, teniendo sus diferencias en la información de los metadatos, los mecanismos de control de calidad utilizados, y las acciones a seguir cuando los datos son considerados sospechosos o erróneos.

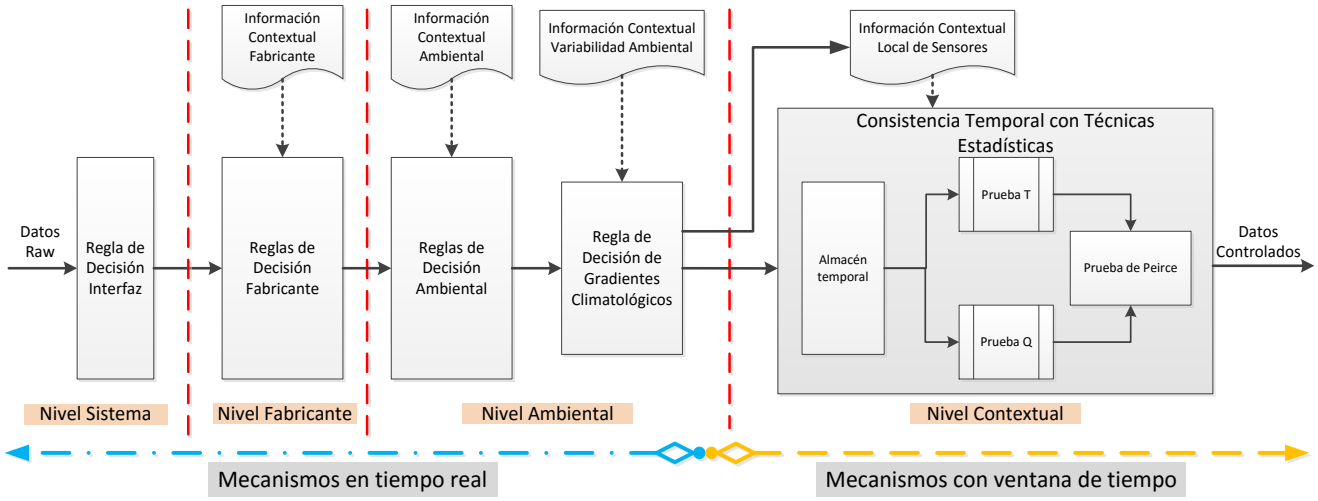


Figura 4-6 Niveles de Control de Calidad.

 Nivel de control del sistema: en este nivel se realiza el control de la medición de los sensores teniendo en cuenta el contexto del sistema en que son tomados. Está destinado principalmente para indicar cualquier mal funcionamiento del sensor, la inestabilidad, el no cumplimiento de valores eléctricos cuando se capturan los datos, problemas de señales débiles cuando exista transferencia de información. Entre los metadatos está el nivel de la batería, la interfaz de comunicación con el sensor para garantizar un flujo continuo de los datos recolectados en una serie de tiempo. Los mecanismos de control de calidad se basan en reglas de decisión sobre las condiciones de los dispositivos ya que pertenece al contexto del nodo. Los valores marcados como erróneos en este nivel de control de calidad puede tener dos alternativas: una alternativa es ignorar los datos erróneos y otra alternativa es eliminar y solicitar una nueva medida, al mismo tiempo se asigna una marca a nivel del sistema de adquisición para reportar una posible alarma de mantenimiento.

- Nivel de control del fabricante: Se realiza el control de un tipo de datos (datos univariados), teniendo en cuenta los parámetros e información del fabricante, se revisan los umbrales de la medición para valores operacionales (Dutta et al. 2014). Los mecanismos de control de calidad utilizados se basan en reglas de decisión basados en rangos operacionales, función umbral y filtros lógicos (Estévez et al. 2011). Los posibles datos erróneos se marcan y se almacenan para determinar las alarmas del sensor y programar planes de calibración y mantenimiento, además se van estructurando los datos a un modelo estándar de control de calidad.

- Nivel de control ambiental: Se realiza el control de un tipo de datos (datos univariados), teniendo en cuenta los parámetros de referencia ambiental)información contextual ambiental y variabilidad climática) que provienen de datos históricos y resultados de análisis realizados por fuentes externas en la zona de estudio; los mecanismos de control de calidad utilizados se basan en reglas de decisión basada en rangos climatológicos máximos y mínimos permitidos en la zona, reglas de decisión de gradientes climatológicos sobre la serie temporal de datos comparada con análisis previos sobre el fenómeno climatológico de estudio. Otras técnicas basadas en patrones de referencia a nivel agroclimatológicos, pruebas de tolerancia en donde se establecen límites superiores o inferiores, pruebas de coherencia interna en cada sensor, pruebas de coherencia temporal (Atzberger 2013). Los posibles datos anómalos se marcan como sospechosos o perdidos, además se generan nuevos datos como la media y la desviación estándar y se determinan valores máximos y mínimos de la serie temporal, llamada información contextual local. Los datos sospechosos se pueden corregir remplazando el dato por la media, interpolar o ignorar el dato anómalo.

- Nivel de control contextual: se realiza el control teniendo presente múltiples instancias de un tipo de datos (datos multivariados), este proceso incluye definir una ventana de tiempo para almacenar los datos temporal, además requiere de datos ordenados y que ya hayan sido filtrados en otros procesos (datos limpios) ya que los mecanismos de control de calidad usan técnicas estadísticas: prueba Q de Dixon, prueba T o ASTM, el criterio de Peirce´s, el criterio de Chauvenet´s. los valores que no pasen la prueba se etiquetan como inconsistentes y los demás como buenos, las alternativas para los datos inconsistentes es ignorar el dato, y otra alternativa es interpolar con la muestra anterior y siguiente en la serie temporal.

4.5 Motor de actualización

La mejora continua del proceso de control de calidad se implementa basada en un aprendizaje por refuerzo que observa y aprende del contexto local, ejecuta una acción orientada a la actualización de parámetros de referencia y trata de seleccionar la mejor alternativa propuesta (ignorar, medir, reemplazar e interpolar), planteadas al final del proceso de control de calidad de los datos, cuando estos son anómalos.

En el enfoque de aprendizaje por refuerzo se tiene un conjunto de estados y eventos, que ejecutan acciones, dentro de este enfoque el método más utilizado en redes de sensores - WSN es el Q-learning, que define un Q-value en el tiempo (t) a partir de un Q (estado, evento, acción) que es actualizado usando una recompensa en el tiempo (t+1). (Figura 4-7).

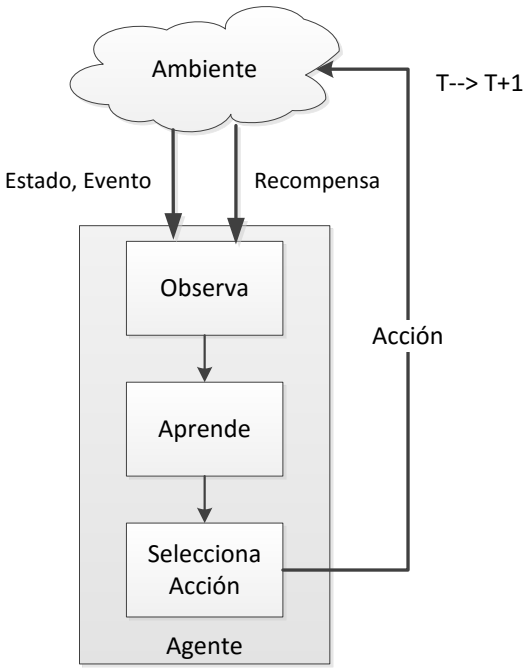


Figura 4-7 Método de aprendizaje por refuerzo Q-learning.

En nuestra propuesta el método de aprendizaje aprende de la información contextual local generada por el mecanismo de control de calidad, representada por valores estadísticos de la media y desviación estándar incremental, al igual que los valores máximos y mínimos de la señal, siempre y cuando los datos son etiquetados como buenos y calculados en la ventana de tiempo, obteniendo un nuevo conjunto de

parámetros de referencia, que debe ser comparado con el conjunto de parámetros de referencia actual por medio de unas reglas de actualización que deciden si se necesita o no actualización de parámetros.

El conjunto de parámetros de referencia se consumen en los mecanismos de control de calidad generando nuevos datos controlados, etiquetados con marcas de calidad que infieren la cantidad de datos anómalos y datos buenos, ofreciendo recompensas, para actualizar el nuevo juego de reglas y las acciones a ejecutar en un nuevo proceso de aprendizaje.

Las recompensas buscan aumentar la precisión y la sensibilidad del mecanismo de control de calidad (una alta precisión hace que el mecanismo recupere resultados más relevantes que irrelevantes y una alta sensibilidad hace que el mecanismo retorne más resultados relevantes), para ello debemos de aumentar el número de datos anómalos clasificados como anómalos (verdaderos positivos - TP) y disminuir el número de falsos positivos (FP) y falsos negativos (FN).

Los estados se basan en el nivel de calidad de los datos, representada como rangos de valores en los cuales se etiquetan los datos con marcas de calidad, un ejemplo de estado es cuando un valor esta por fuera del rango aceptable del sensor se etiqueta como erróneo (estado_fuera), si está dentro del rango del fabricante pero fuera del rango climatológico se etiqueta como dudoso (estado_fabricante), si está dentro del rango climatológico pero el dato no cambia con el tiempo se etiqueta como perdido (estado_constante) y si la variación entre el dato anterior d(t) y el dato siguiente d(t+1) supera un valor predefinido se etiqueta como inconsistente (estado_inconsistente) y si el dato supera satisfactoriamente todos los mecanismos de control de calidad se etiqueta como bueno (estado_bueno).

Los eventos son las muestras capturadas por los sensores cada intervalo de tiempo (datos *raw*) y las acciones se orientan a la modificación de un parámetro de referencia que ayuda a que se genere un cambio de estado cuando llega un nuevo dato.

En conclusión tenemos un Q (nivel de calidad, datos *raw*, modificar parámetros) ➔ recompensa (verdaderos positivos, falsos negativos y falsos positivos). En la Figura 4-8 se detalla este proceso.

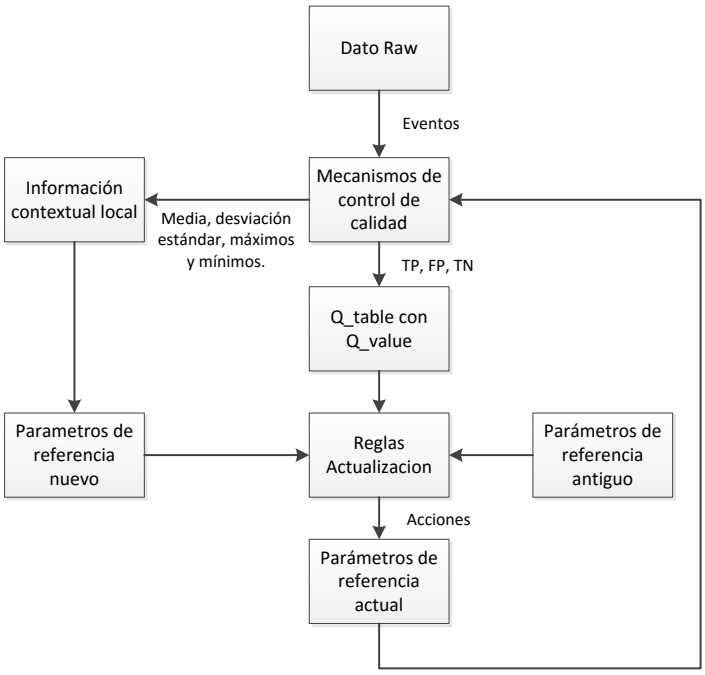


Figura 4-8 Motor de actualización de parámetros con aprendizaje por refuerzo.

La regla de actualización basada en la desviación estándar de la ventana de tiempo, nos da una indicación de que tanto se aleja una muestra con respecto a otra, para ello tomamos el valor de dos veces la desviación estándar de la ventana de tiempo anterior para ser usado como gradiente de variabilidad.

$$Si\ \left((\sigma_{v(t)} < \sigma_{v(t+1)}) and\ \left(\sigma_{v(t+1)} < 2\sigma_{v(t)}\right)\right)\ entonces\ \lambda_{max} = 2\sigma_{v(t+1)}$$

En donde $\sigma_{v(t)}$ es la desviación estándar de la ventana en el tiempo t y $\sigma_{v(t+1)}$, la desviación estándar en la ventana de tiempo t+1, λ_{max} es el límite de variabilidad máximo entre una muestra x(t) y la muestra x(t+1), mostrado en el algoritmo 4 (regla de decisión de gradientes climatológicos) en la prueba pendiente y tendencia. En pruebas simuladas tenemos un incremento en la sensibilidad y en la precisión. (Tabla 4-1).

Escenario	Sensibilidad	Especificidad	Precisión	Falsas Alarmas
Discontinua con ruido gaussiano(4,3) $\lambda_{max}\ fija$	89,1%	84,4%	65,5%	15,6%
Discontinua con ruido gaussiano(4,3), $\lambda_{max}\ variable$	98,0%	83,3%	66,1%	16,7%

Tabla 4-1 Simulación del motor de actualización en la prueba de tendencia.

4.6 Salida control de calidad

En la salida de datos de la capa de control de calidad, haremos uso de los estándares interoperables propuestos por el SWE, tales como: a) observaciones y medidas (O&M) define el esquema conceptual de una observación, entendida como una acción cuyo resultado es el valor estimado de una propiedad de un aspecto de interés en un tiempo específico y usando un procedimiento especifico; b) servicio de observación del sensor (SOS) servicio creado para proporcionar datos de observaciones, como resultado del modelo descrito en el estándar O&M; c) SensorML estándar que describe y estructura la información procedimental del sensor, usada en la comunicación entre O&M y SOS, para recuperación de observaciones. En la Figura 4-9 se puede observar el proceso de registrar un sensor en el servicio SOS para recuperar las observaciones o datos de interés de un fenómeno.

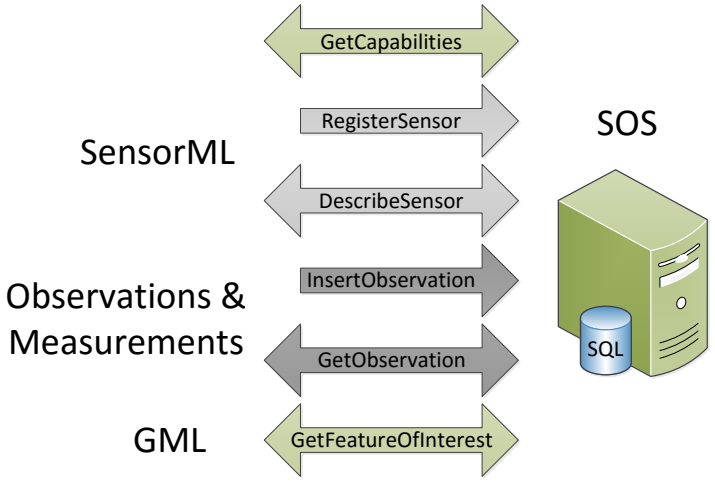


Figura 4-9 Interfaces interoperables con estándares SWE de OGC.

La interacción entre estándares para recuperar información de un sensor, comienza especificando las características procedimentales del sensor por medio de SensorML, a continuación se debe asociar un sensor a un servicio SOS (*RegisterSensor*), los detalles del sensor lo puede leer el servicio SOS (*DescribeSensor*), ya vincula el sensor, se procede a recuperar una observación por parte del servicio SOS (*GetObservation*), o también el sensor puede enviar una observación para ingresar nuevos datos (*InsertObservation*), si es necesario el servicio SOS recupera los metadatos del servicio (*GetCapabilities*), o el aspecto de interés asociado a una observación (*GetFeatureOfInterest*).

SensorML es usado para describir los metadatos del sensor, incluyendo una identificación, clasificación, entradas, salidas, parámetros y características tales como descripción espacial y temporal. (Figura 4-10).

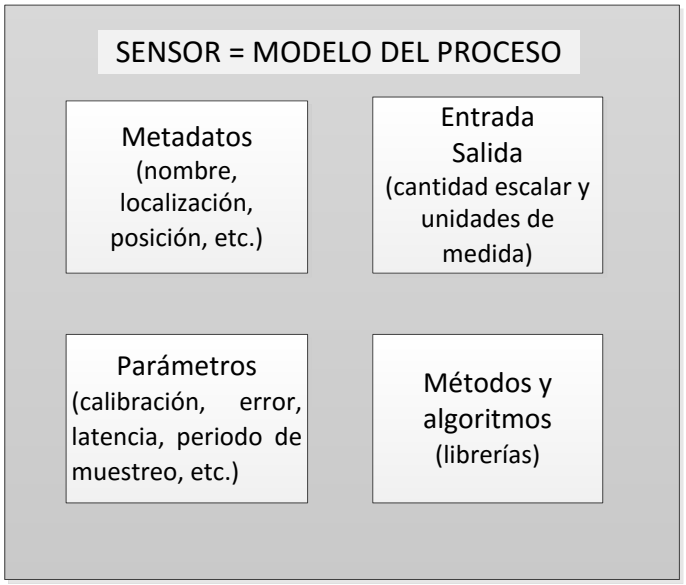


Figura 4-10 Modelado de un sensor con SensorML.

El O&M define una observación como el acto de observar cierto fenómeno, tiene en cuenta un proceso (sensor), la propiedad observada que es la descripción del fenómeno (temperatura del aire), el fenómeno presenta un aspecto de interés, que es una entidad del mundo real que es objeto de observación (monitoreo ambiental de un cultivo), la observación proporciona un valor que es la propiedad observada en un instante de tiempo (24112015 08:20:05.152), este puede ser de cualquier tipo, un valor numérico (19°C) o un valor nominal (caliente). Además el valor tiene asociado una calidad (intervalo de error de 0.03°C). (Figura 4-11).

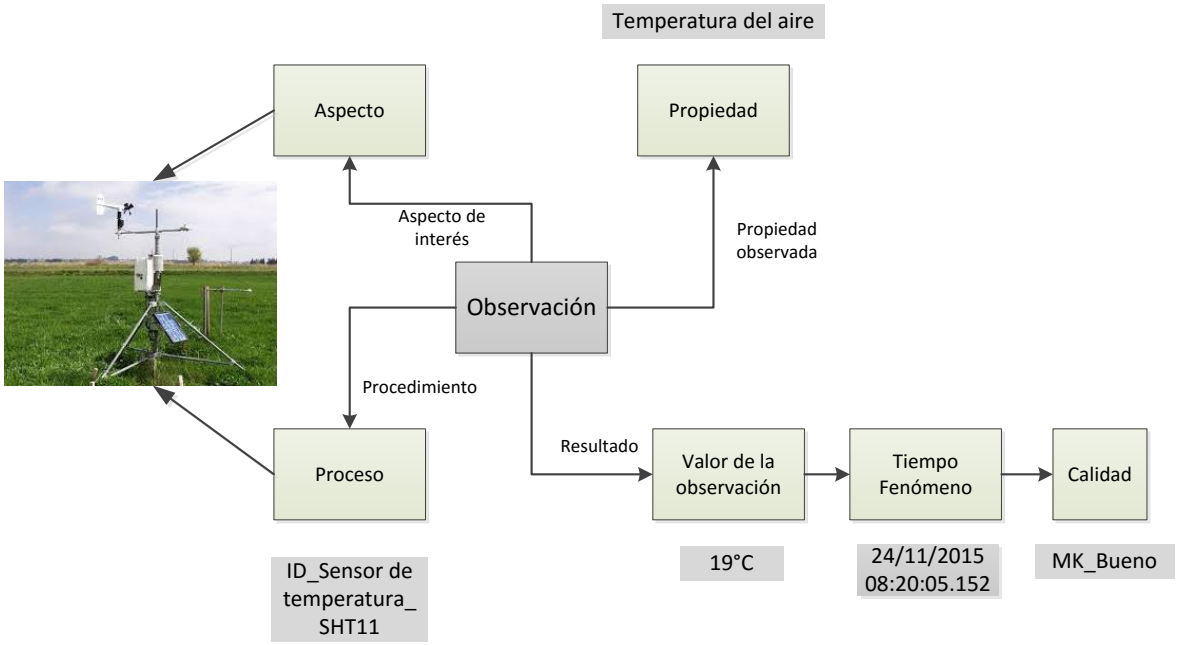


Figura 4-11 Sistema propuesto de observación O&M.

La estructura interna del módulo y su interacción con un servidor se ve en la Figura 4-12, dentro del sistema de adquisición se diseña un modelo de datos nativo, que permite almacenar cierta información relacionada con los datos capturados por los sensores, el tiempo de captura, la marca de calidad asignada durante el proceso de control de calidad, el identificador del dato y los valores originales de los datos; este modelo de datos nativo se debe convertir a un modelo de datos XML que cubra los requerimientos del estándar O&M, para ser interoperable y enviar los resultados de las observaciones a un servidor remoto que ejecuta el servicio de observación del sensor SOS.

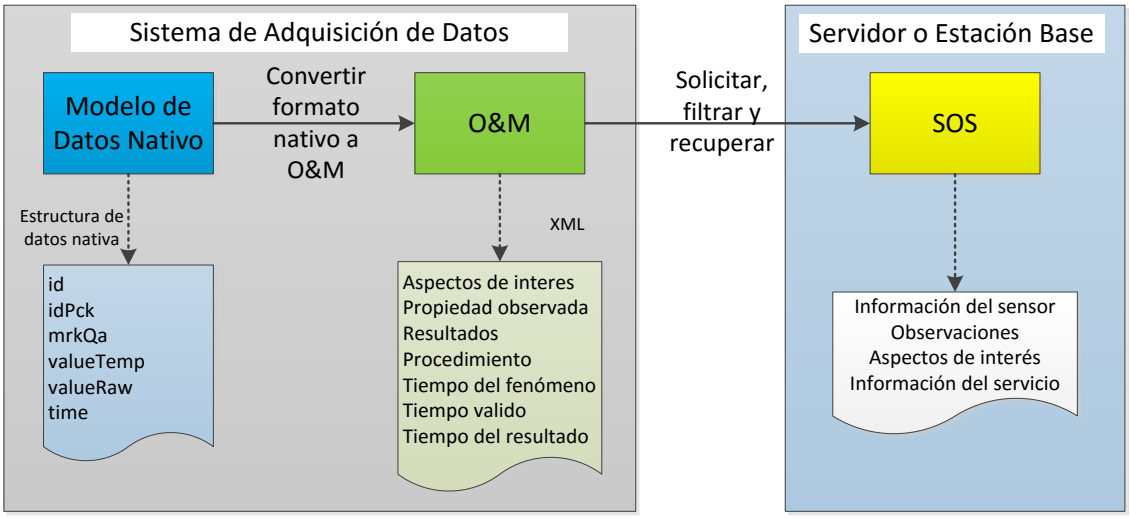


Figura 4-12 Componente de salida de datos e interacción con un servidor SOS.

En nuestra propuesta, solo se estructura un modelo O&M, pero no se realiza todo el proceso de recuperación de la información del sensor y sus observaciones por medio de SOS, ya que esta por fuera del alcance de este trabajo.

4.7 Resumen

La arquitectura propuesta presenta el diseño de un motor de control de calidad que incluyen los mecanismos definidos en el capítulo 3, al mismo tiempo la inclusión de un motor de actualización basado en un enfoque de aprendizaje por reforzamiento que busca actualizar parámetros de referencia con el uso de información contextual local generada durante la captura de los datos de los sensores, por otro lado la forma como se formatea la salida de los datos del procesos de control de calidad a los estándares SWE de la OGC, entre ellos observaciones y medidas (O&M) y el servicio de observación del sensor (SOS). La información contextual global se modela por medio de SensorML, que refleja las características del sensor y del sitio de observación.

Al final la concepción del modelo desde el punto de vista físico, conceptual, interoperable y con aprendizaje, se traslada a unos componentes claves que permiten realizar el control de calidad de datos por medio de un modelo lógico y unos niveles de control que conducen los datos *RAW* de los sensores a datos controlados y compartidos con los centros de datos, al mismo tiempo generando datos procesados útiles para mejorar continuamente el mecanismo de control de calidad de datos agroclimatológicos.

Capítulo 5

Evaluación de la propuesta

En este capítulo se implementa la arquitectura del sistema propuesto junto con los mecanismos de calidad de datos seleccionados, la cual es objeto de evaluación en un ambiente real. Se selecciona la plataforma *Waspmote Plug & Sense* y la *board* de agricultura inteligente (*Smart* agricultura PRO) de la empresa Libelium (Libelium 2013), cuyo ambiente de programación *Waspmote* PRO IDE está orientado a objetos en lenguaje de programación C++.

El sitio de observación o estudio es la ciudad de Popayán localizada en el departamento del Cauca - Colombia. Al inicio se modela cada uno de los componentes de la arquitectura propuesta (la información contextual y los mecanismos de control de calidad) en clases, se crea una librería para ser adaptable al entorno de desarrollo, se construye la lógica del programa principal con base en los niveles de control de calidad definidos anteriormente y se evalúa todo el sistema por los escenarios descritos en el capítulo 3, etiquetando los datos de salidas con marcas de calidad.

5.1 Implementación

La implementación se realiza sobre WSN de libelium (*Waspmote Plug & Sense*) y unas tarjetas para cada aplicación (*Smart* agricultura), de la información técnica del sensor de temperatura (SHT75) se construye la información contextual del fabricante del sensor y del sitio de observación (Popayán – Cauca) se extrae información contextual de temperatura y humedad de portales climatológicos. Y por último se diseña cada uno de los componentes de la arquitectura propuesta en el ambiente de programación *Waspmote* PRO IDE.

5.1.1 WSN libelium

Libelium es una empresa dedicada a diseñar módulos y tarjetas relacionadas con redes de sensores inalámbricas – WSN, van desde el diseño de los sensores, el procesamiento o microcontrolador de bajo consumo, hasta propuesta de tecnologías inalámbricas (802.15.4, *Xbee, Bluetooth, WiFi, GPRS/3G, 6LowPan, LoRa, Sigfox*, etc.) esto en cuanto a hardware, además cuentan con un entorno de desarrollo integrado *IDE* llamado Waspmote Pro *API*, que permite la configuración y programación de los nodos sensores, al mismo tiempo presentan una solución de estación base llamada Meshlium, que permite capturar información de los nodos sensores, con diferentes tecnologías, buscando un almacenamiento temporal para ser enviado posteriormente a servidores en la nube por medio de otras tecnologías.

De esta manera con el módulo Waspmote, los datos capturados por lo sensores se pueden visualizar en la nube, a través de diferentes plataformas de computación (cloud computing), en conclusión Waspmote logra conectar sensores a la nube. En la Figura 5-1 se puede observar el diseño de una *WSN* con Waspmote y Meshlium.

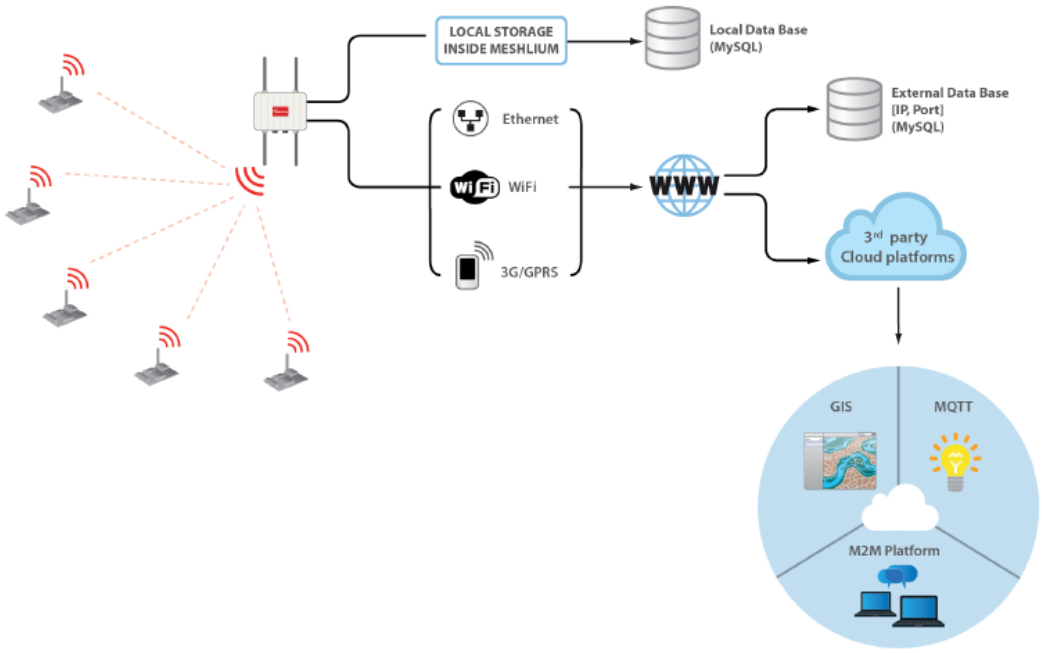


Figura 5-1 WSN de Libelium con Waspmote y Meshlium.

Fuente: Libelium (*Waspmote Plug & Sense.* (Libelium 2013)

El módulo utilizado para agricultura inteligente (Figura 5-2), cuenta con sensores de temperatura, humedad, precipitación, velocidad y dirección de viento, presión atmosférica y radiación solar, dentro de las aplicaciones se tiene agricultura de precisión, sistemas de riego, estaciones climatológicas, etc.

Figura 5-2 Modulo Waspmote Plug & Sense para agricultura inteligente.

Fuente: Libelium (*Waspmote Plug & Sense* (Libelium 2013)

El ambiente de desarrollo integrado (*Waspmote* Pro API) permite la generación automática del código y la programación de los nodos sensores por dos vías: USB y programación por el aire. (Figura 5-3).

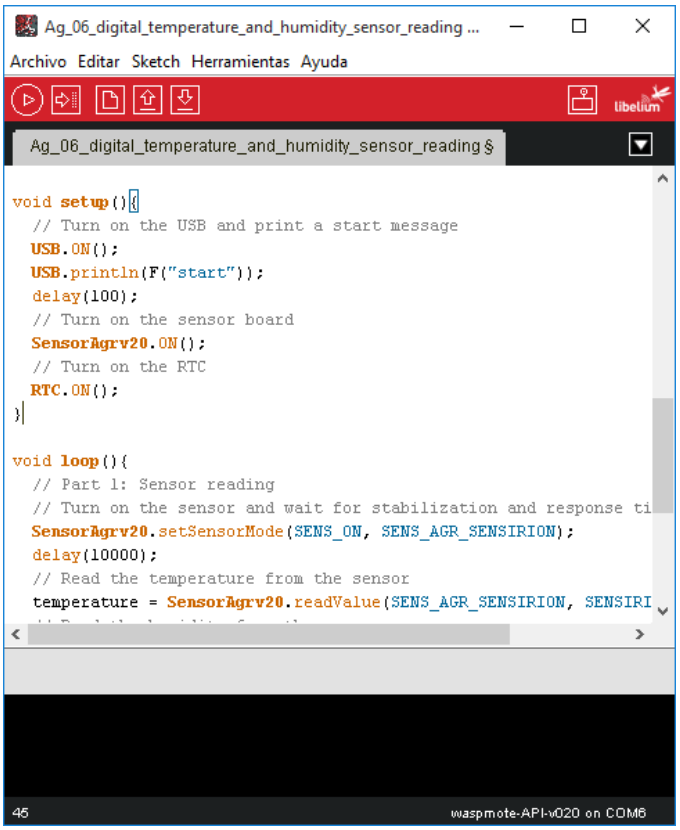


Figura 5-3 Waspmote PRO IDE.

Este entorno de desarrollo IDE, presenta un conjunto de herramientas que hace posible realizar un código entre las cuales se tienen las librerías (sensores y tecnologías de comunicación) así mismo unos ejemplos agrupados como: general, sensores, comunicación y combinados, esto con el fin de facilitar las labores de programación ya que presenta un aspecto muy similar a Arduino IDE[1].

5.1.2 Sensor de temperatura

Sensor de temperatura SHT75 del fabricante Sensirion, toma medidas de temperatura y humedad, es un integrado encapsulado para manejo en exteriores con una resolución de 14 bits, la luz directa no le afecta, no hay transferencia de calor de los pines externos al sensor, la exactitud es de +/- 0.4°C en un rango de 0°C a 70°C que es lo necesario para las condiciones ambientales de la zona de estudio.

[1] http://www.arduino.cc

Otro dato importante del sensor es el tiempo de respuesta ya que determina el intervalo de tiempo requerido para actualizar al nuevo valor correcto a la salida, con un grado de tolerancia aceptable, en este caso del 63%. (Figura 5-4).

Sensor specifications (SHT75)

Power supply: 2.4 ~ 5.5V
Minimum consumption (sleep): 2µW
Consumption (measurement): 3mW
Average consumption: 90µW
Communication: Digital (two wire interface)
Storage temperature: 10 ~ 50ºC (0 ~ 80ºC maximum)
Storage humidity: 20 ~ 60%RH

Temperature:
Measurement range: -40ºC ~ +123.8ºC
Resolution: 0.04ºC (minimum), 0.01ºC (typical)
Accuracy: ±0.4ºC (range 0ºC ~ +70ºC), ±4ºC (range -40 ~ +125ºC)
Repeatability: ±0.1ºC
Response time (minimum): 5 seconds (63% of the response)
Response time (maximum): 30 seconds (63% of the response)

Figura 5-4 Sensor de temperatura y humedad SHT75.

Esta información determina los metadatos del fabricante que son insertados en el momento de la configuración inicial, por ser estáticos no necesitan de actualización, al menos que se cambie el sensor por otra referencia. Esta información contextual externa es almacenada en una memoria estática y no borrable, además de estar estructurada (Figura 5-5) para ser utilizada por el sistema de control de calidad.

MetaTemp
-name : char -type : int -volt : double -amp : double -minHum : double -maxHum : double -minTemp : double -maxTemp : double -rate : int -accuracy : int
+getName() : char +setName() +getType() : int +setType() +getVolt() : double +setVolt() +getMinTemp() : double +setMinTemp() +getMaxTemp() : double +setMaxTemp() +getRate() : double +setRate()

Figura 5-5 Estructura de la clase metadatos de sensor de temperatura.

Existen otras características que nos permite gestionar fallas y alarmas del sensor entre ellas se tienen: los datos de calibración se pueden ejecutar nuevamente por medio de una instrucción, logrando así el mantenimiento del sensor, otro elemento con que viene incluido el chip es con un calentador que permite registrar como temperatura del sensor un valor de 5 - 10 °C permitiendo hacer un análisis con la humedad relativa que es capturada del ambiente, sumado a estas características se notifica de un nivel de voltaje menor a 2.47 voltios.

Dicha información representa en unos casos acciones cuando los datos salen anómalos y requieren un mantenimiento del sensor, en otros casos para inducir fallas y evaluar el mecanismo de control de calidad propuesto.

5.1.3 Sitio de observación

La temperatura promedio es de 18.7 ºC. Al medio día la temperatura máxima media oscila entre 24 y 25ºC. En la madrugada la temperatura mínima está entre 12 y 14ºC. El sol brilla cerca de 4 horas diarias en los meses lluviosos; en los meses secos, la insolación es levemente inferior a 6 horas diarias/día. (Figura 5-6).

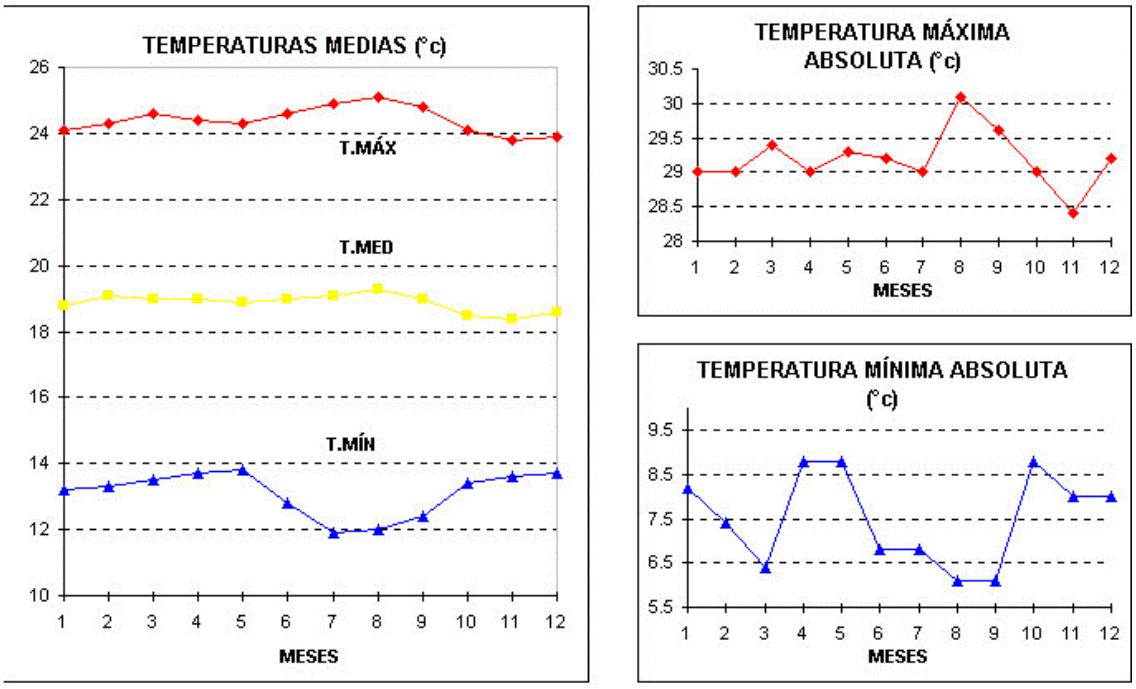


Figura 5-6 Información Climatológica del sitio de observación.

Fuente http://bart.ideam.gov.co/cliciu/popa/temperatura.htm

La localización geográfica de Popayán se encuentra a 76.61 grados de longitud oeste, 2.44 grados de latitud norte, su humedad relativa promedio es de 74%. La implementación de esta clase se puede observar en la Figura 5-7. Donde se tienen los metadatos ambientales (MetaAmb) y las coordenadas geográficas (GeoLocac).

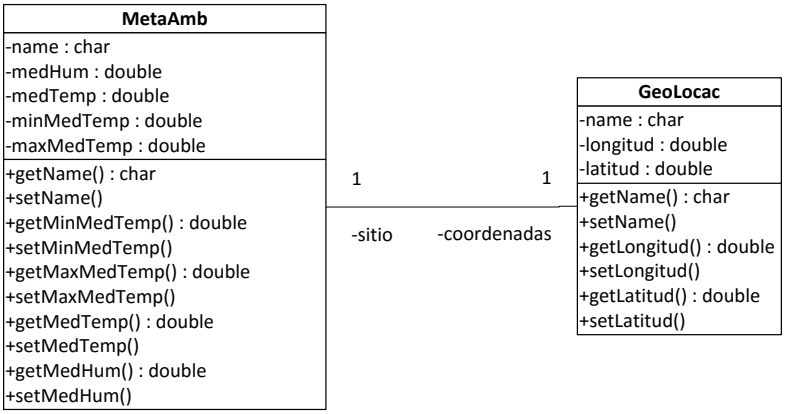


Figura 5-7 Estructura de la clase metadatos del sitio de observación.

5.2 Implementación del mecanismo

El sistema de control de calidad se decide implementar como una librería que puede ser importada por el entorno de desarrollo integrado Waspmote Pro, ya que el ambiente de programación es un lenguaje orientado a objetos (C++), se estructuran diferentes clases entre ellas la clase Quality (modela los mecanismos de control de calidad y genera la información contextual local), la clase DataTemp (implementa el modelo de datos nativo), la clase MetaTemp (almacena la información contextual que proviene de fuentes externas en este caso las características del sensor de temperatura utilizado SHT75). Y la clase MetaAmb (modela la información contextual del sitio de observación en nuestro caso la ciudad de Popayán Cauca Colombia). La clase Sensor se refiere a las lecturas de temperaturas realizadas, la cual es la fuente original de los datos, que se muestrea cada cierto tiempo. La clase *Failed* genera las fallas necesarias para evaluar el sistema propuesto en un ambiente real. En la Figura 5-8 se muestra un diagrama de clases con los métodos y atributos utilizados.

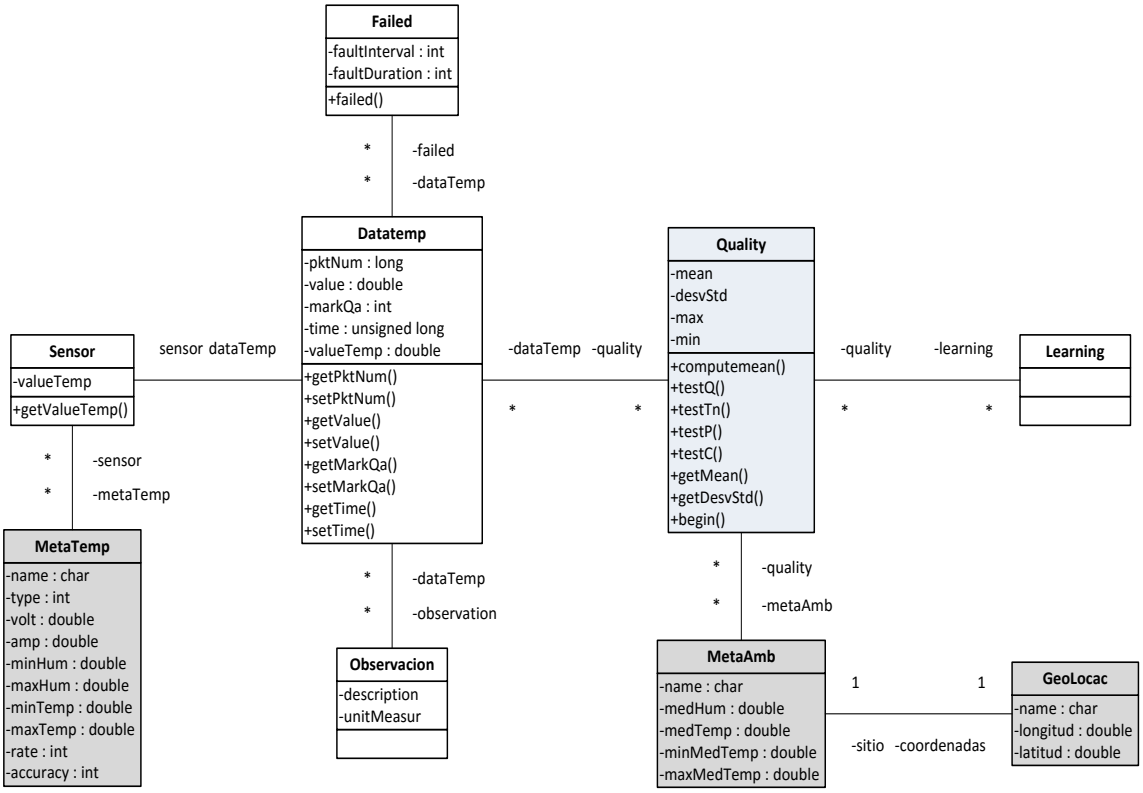


Figura 5-8 Diagrama de clases del sistema en el ambiente Waspmote PRO IDE.

La librería se puede importar desde el ambiente de desarrollo (Figura 5-9), es completamente funcional, con los atributos de edición e interfaz gráfica de las demás librerías, además se puede utilizar en otros escenarios y con otros sensores. Sumado a esto se crearon varios ejemplos que están a disposición de los usuarios de Libelium.

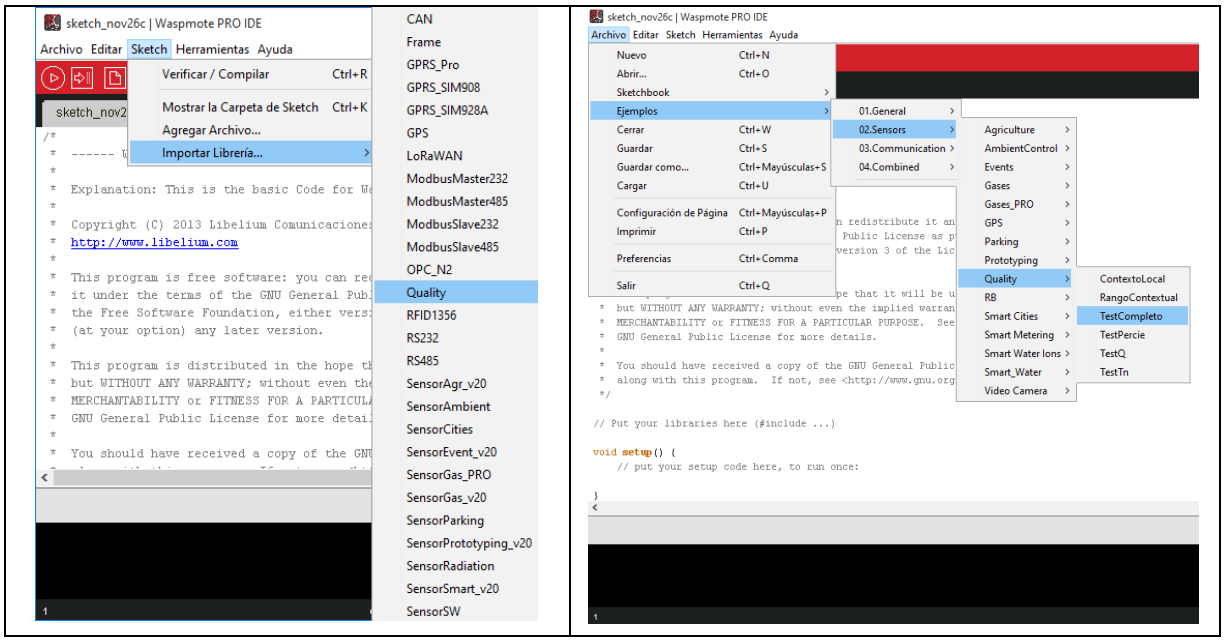


Figura 5-9 Librería Quality y ejemplos control de calidad en Waspmote PRO IDE.

En la Figura 5-10 se muestra la implementación del mecanismo de control de calidad haciendo uso de las clases definidas anteriormente, en el lado izquierdo se observan los niveles definidos por reglas contextuales, ya sean con información del fabricante que es la primera que se evalúa y después con valores climatológicos del sitio de observación, al terminar esta etapa con valores que han superado esta prueba se realiza la prueba de variabilidad entre la muestra anterior y la actual en la serie temporal. Al mismo tiempo se genera y almacena información contextual local (media, desviación estándar, máximos y mínimos).

En el lado derecho de la figura se implementa las técnicas estadística que almacena temporalmente los datos definidos por una ventana de tiempo, aquí se realiza la prueba T y se estructura la trama de datos para ser enviada por el puerto serial.

Cabe anotar que el diseño esta optimizado en energía, activando cada componente hardware cuando se necesite. Por ello se activa primero el sensor y se captura la muestra, posteriormente se apaga y comienza el proceso de control de calidad, al finalizar se activa el módulo de comunicación para enviar la trama al exterior y por último se duerme todo el nodo sensor para despertarse después de un tiempo establecido, en nuestro caso 20 segundos.

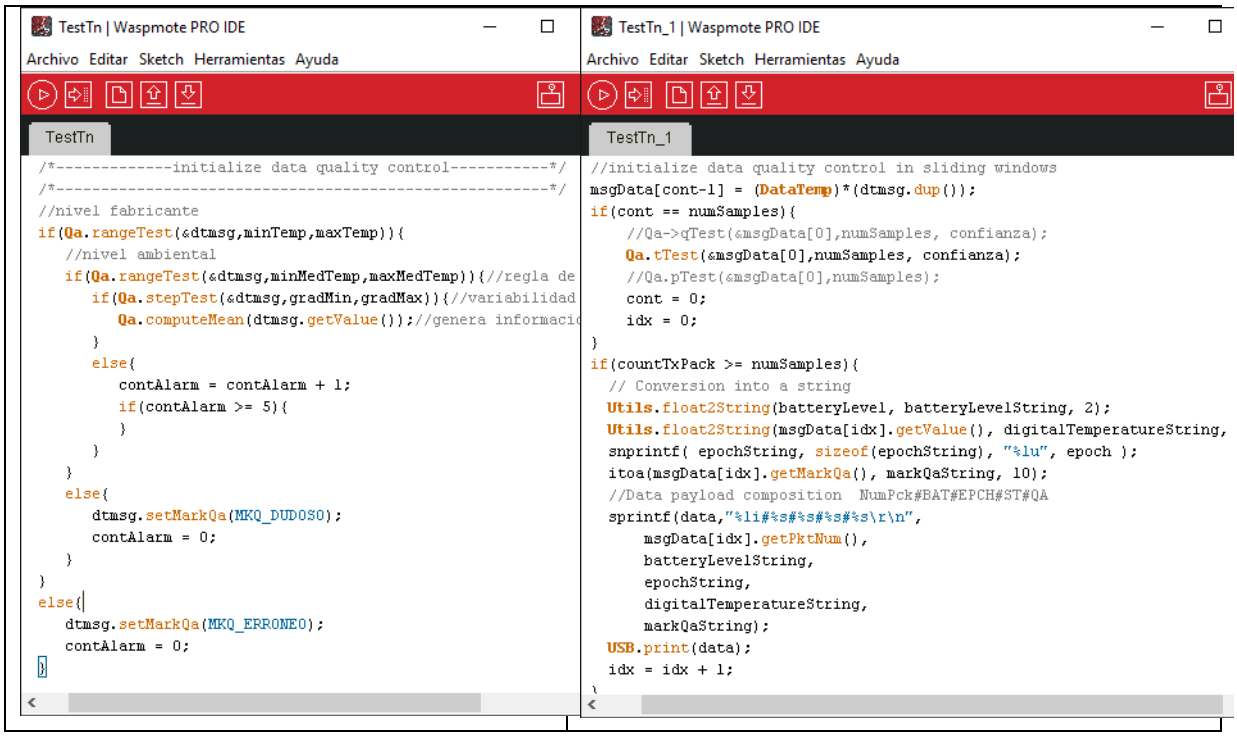


Figura 5-10 Implementación de mecanismos de control de calidad.

Se estructura una trama de datos para ser enviada por el puerto serial donde se establece una identificador del paquete según aparición en la seria temporal de datos, enseguida se inserta el valor de voltaje de la batería, sumado a esto se inserta el tiempo en formato *epoch*, el valor de la muestra del dato tomado del sensor de temperatura y por último la etiqueta de calidad asignada después de realizado el proceso de control de calidad de los datos.

Id_Dato	Voltaje_Batería	Tiempo_Epoch	Valor Temperatura	Marca Calidad

En la Figura 5-11 se observa la llegada de datos por el puerto serial haciendo uso de la utilidad de monitor serial presente en el ambiente de desarrollo Waspmote PRO IDE.

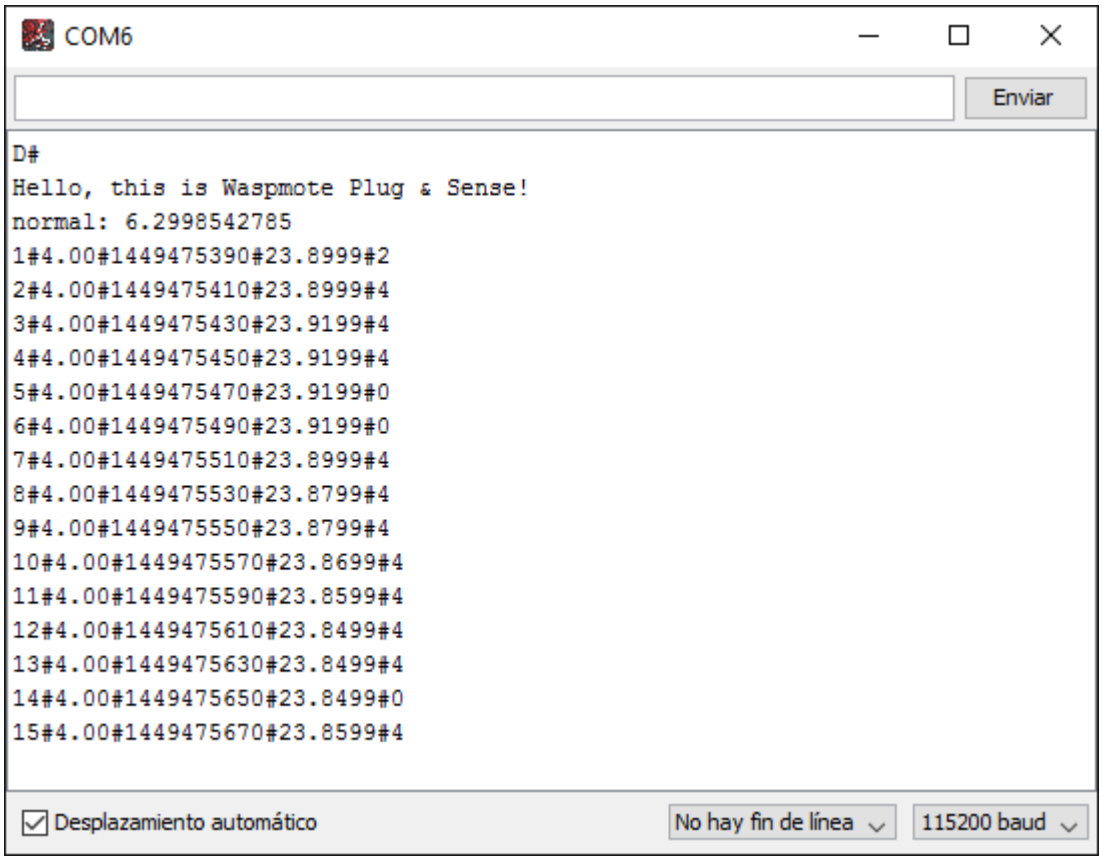


Figura 5-11 Tramas de temperatura enviadas con marcas de calidad.

5.3 Escenario de evaluación

La propuesta es insertar fallas de datos (valores atípicos, ruido excesivo, fallas contantes y valores fuera del rango) durante la captura de información del sensor de temperatura para que el mecanismo clasifique los datos como buenos o anómalos, y por medio de las métricas de evaluación analizar su resultados, entre los escenarios a evaluar tenemos 1) Fallas continuas se produce durante todo el experimento. 2) Fallas discontinuas ocurren a intervalos regulares de tiempo; estas fallas discontinuas se caracterizan por dos parámetros: la duración de la falla establecida en 100 segundos, y el número total de apariciones durante el experimento cada 300 segundos. (Figura 5-12).

Sin Fallas

1. Ideal sin ningún tipo de falla.

Fallas constantes (*Struck-at*)

2. Falla Continua con valor constante de 22.56°C.

3. Falla Discontinua con valor constante igual a la muestra anterior.

Fallas con ruido gaussiano (*Outliers* y *Spike*)

4. Falla Continua con ruido gaussiano (4, 3.0) °C.

5. Falla Discontinua con ruido gaussiano (4, 3) °C.

Fallas con valores fuera de rango (*Out-of-range*)

6. Falla Discontinua con valores fuera de rango 34 °C.+ función Normal (4,3) °C.

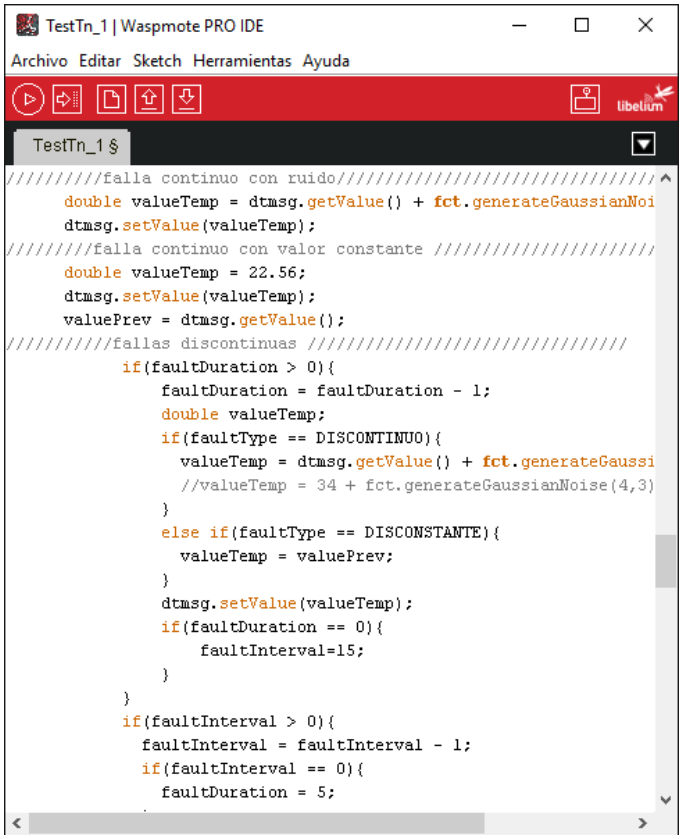
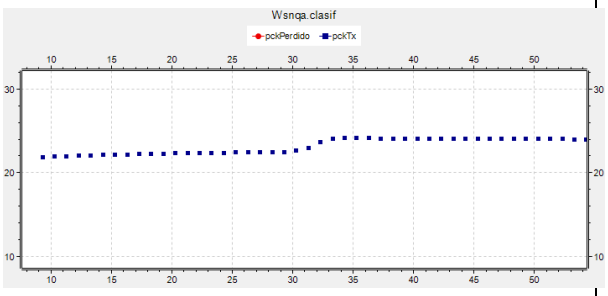
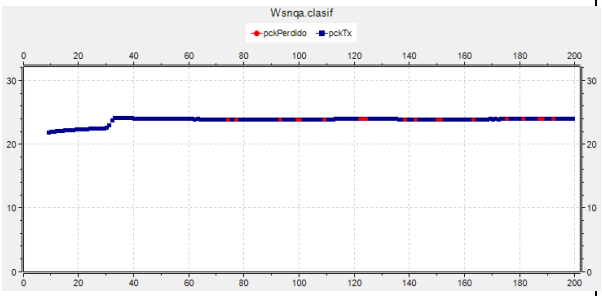
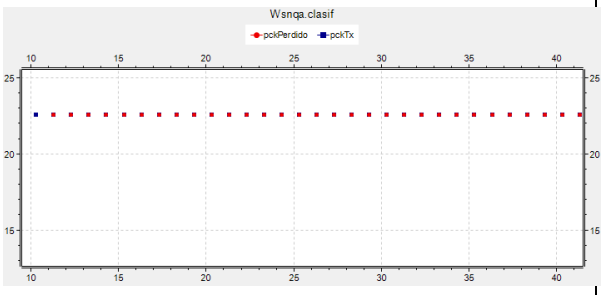
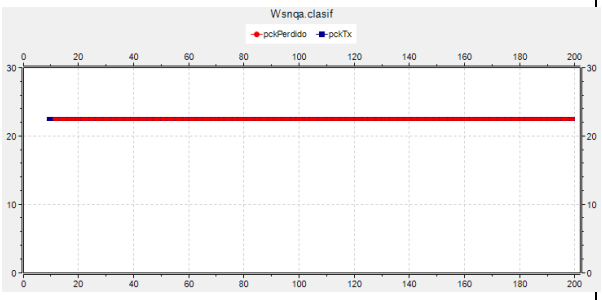
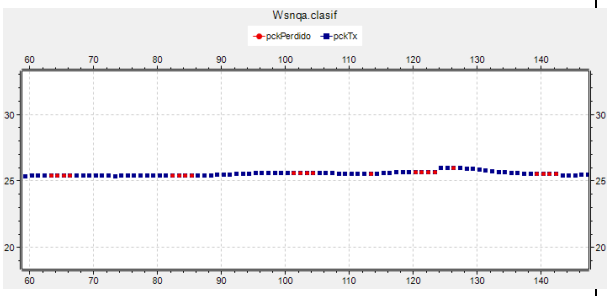
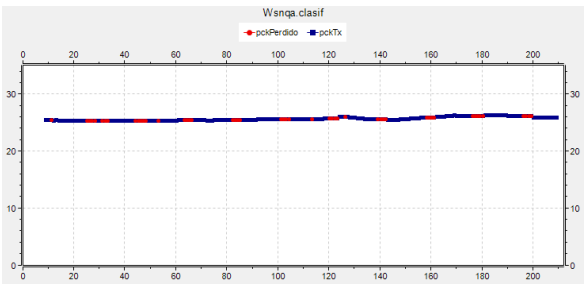


Figura 5-12 Implementación de fallas.

5.4 Resultados de efectividad

A continuación se presenta el análisis de los resultados de la arquitectura implementada con los mecanismos de control de calidad ante los diferentes escenarios de fallas en un ambiente real (plataforma Waspmote de Libelium). En la Figura 5-13 se muestra a la derecha las gráficas del conjunto de datos seleccionado con 200 muestras para cada uno de los escenarios de prueba y a la izquierda una imagen ampliada del tipo de falla insertada en la fuente original de datos y etiquetada por medio de marcas de calidad.

Ideal sin ningún tipo de falla	
	
Falla Continua con valor constante de 22.56 °C.	
	
Falla Discontinua con valor constante	
	

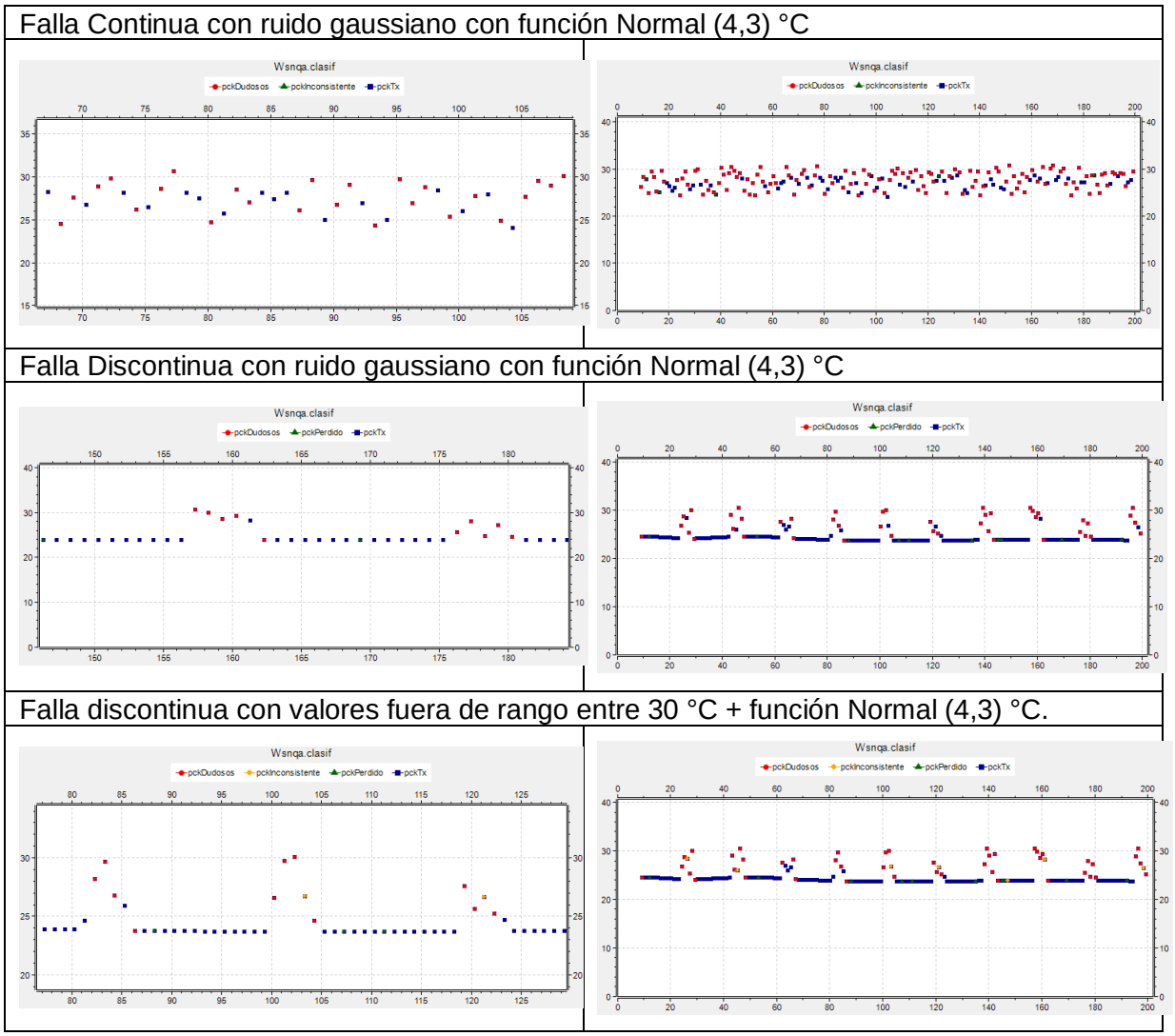


Figura 5-13 Resultados evaluación de todo el sistema propuesto.

En la Tabla 5-1 se muestra los resultados del proceso evaluación en un escenario real evaluando cada una de las fallas propuestas, aquí se evidencia que en un escenario de falla con un valor constante ya sea continuo o a intervalos de tiempo definidos con anterioridad (discontinuo).

La etiqueta MKQ_PERDIDO representa el mayor número de datos anómalos cuando el ruido presente es constante durante una intervalo de tiempo considerable, como se evidencia en el escenario de falla continua (198) o discontinua (55) de una población de 200 muestras, controlando fallas constantes.

En cuanto a fallas de ruido y valores atípicos la etiqueta MKQ_DUDOSO, presenta el mayor número de datos anómalos, siendo el mecanismo de control de calidad un rango contextual basado en parámetros climatológicos, aquí la técnica estadística no presenta mucho trabajo (MKQ_INCONSISTENTE con valores de 6 detecciones) pero es necesaria para fortalecer el control de calidad de los datos.

El algoritmo de técnica estadística es muy robusto para detectar anomalías, es por ello que se utiliza en una capa interior del proceso, para que sea más efectivo y no se contamine con valores fuera de rango o fallas constantes.

Por ultimo en las fallas con valores fuera del rango, se detectan datos anómalos por medio del mecanismo de funciones umbral, en este caso los parámetros o rangos climatológicos de la zona de estudio son los principales en la detección de anomalías, además se etiquetan como MKQ_DUDOSO (51 detecciones).

Como resumen del análisis, los diferentes mecanismos de control de calidad ayudan en la detección de un tipo de falla o anomalías, de la combinación de ellos se obtienen mejores resultados en la detección, además el etiquetado del proceso ayuda a visualizar e identificar el tipo de falla presentada y así tomar las respectivas acciones a ser corregidas en un proceso posterior.

Escenario	TX	MKQ BUENO	MKQ INCONSISTENTE	MKQ DUDOSO	MKQ ERRONEO	MKQ PERDIDO
Ideal	200	180	0	0	0	20
Continua con valor constante	200	1	0	1	0	198
Discontinua con valor constante (2500)	210	153	1	1	0	55
Continua con ruido gaussiano Normal(4,3)	200	60	6	134	0	0
Discontinua con ruido gaussiano Normal(4,3)	200	136	6	46	0	12
Discontinua con valores fuera de rango	200	125	0	51	0	24

Tabla 5-1 Resultado etiquetados de los escenarios de falla propuestos.

En la Tabla 5-2 se puede observar que la detección de anomalías constantes en cuanto a sensibilidad es bastante alta (99%) ya que el mecanismo de control de calidad utilizado se basa en discriminar valores que no presentan mucha variabilidad entre muestras y se reportan en la gran mayoría de datos perdidos.

Por otra parte la detección de valores fuera de rango presenta una sensibilidad alta (94%), ya que utiliza técnicas basadas en reglas contextuales del ambiente o sitio de observación que limita la presencia de anomalías fuera de este rango.

Las fallas discontinuas con una sensibilidad del 88% cuando hay ruido definido con una función normal de media 4 y desviación estándar 3. Por último la de menor sensibilidad es un ruido constante con un 60% de sensibilidad.

En cuanto a especificidad en todos los escenarios se evidencia del 81.3% al 88.2%, demostrando que nuestro mecanismo de control de calidad detecta la ausencia de anomalías en datos buenos o normales.

En cuanto a precisión, el mecanismo propuesto tiene mayor precisión cuando las anomalías son continuas, ya sea con fallas constantes (99.5%) y con ruido gaussiano (85.7%). Para los escenarios con fallas discontinuas la precisión baja hasta el 64.9% en promedio, esto se debe al uso de la ventana de tiempo en el mecanismo de control y la forma como se planteó el escenario de fallas ya que al ingresar las fallas discontinuas se genera una serie temporal de valores que entre ellos son normales o buenos, pero comparada con valores buenos en otra ventana de tiempo da como resultado anomalías. La precisión de falla discontinua con ruido gaussiano mejoro a un 68%.

La relación de falsas alarmas indica que en promedio un 13% de los datos buenos son etiquetados como anómalos, sobretodo en el escenario discontinuo. Para el escenario continuo no se presenta dicha relación ya que consideramos que toda la muestra está contaminada con valores anómalos.

Escenario	Sensibilidad	Especificidad	Precisión	Falsas Alarmas
Continua con valor constante	99,0%		99,5%	
Discontinua con valor constante (200)	92,5%	88,2%	64,9%	11,8%
Continua con ruido gaussiano Normal(4,3)	60,0%		85,7%	
Discontinua con ruido gaussiano Normal(4,3)	88,0%	86,7%	68,8%	13,3%
Discontinua con valores fuera de rango	94,0%	81,3%	62,7%	18,7%

Tabla 5-2 Resultados de evaluación: sensibilidad, especificidad, precisión y falsas alarmas.

5.5 Discusión

En conclusión el mecanismo de control de calidad propuesto es efectivo ya que presenta una aceptable sensibilidad con un promedio del 88% y una especificidad con el 86%.

Si evaluamos la precisión con un promedio de 64% y la sensibilidad con un promedio de 88%, concluimos que nuestro mecanismo de control de calidad clasifica de manera aceptable datos anómalos y datos buenos.

5.6 Resumen

En este capítulo se realizó la implementación del mecanismo de control de calidad, siguiendo los componentes de la arquitectura propuesta, sobre el ambiente de desarrollo Waspmote PRO IDE, el modelamiento de la información siguió los estándares SWE de la OGC, por medio de clases. La evaluación completa de la arquitectura y los mecanismos siguieron las métricas vistas en el capítulo 3.

Capítulo 6

Conclusiones y trabajos futuros

Este capítulo proporciona las conclusiones de este trabajo, así como una discusión sobre los posibles trabajos futuros que se derivan de esta propuesta de tesis o de algunos apartados que están por fuera del alcance del sistema de control de calidad de los datos agroclimatológicos para agricultura de precisión.

6.1 Conclusiones

El estudio de las diferentes técnicas de detección de anomalías en un escenario como la WSN permite establecer ciertos criterios a tener en cuenta en la aplicación de dichas técnicas en un escenario como el propuesto, donde estas técnicas se deben realizar localmente, subsanando ciertas limitaciones de hardware y de los mecanismos de control de calidad expuestos en la teoría con la ayuda de información de contexto global, provenientes de fuentes externas como el fabricante y sitio de observación, como también la información generada localmente proveniente de los datos capturados por los sensores.

La evaluación de los mecanismos propuestos con la ayuda de una herramienta de simulación como OMNET y de un conjunto de datos, nos da flexibilidad en probar repetidas veces y con diferentes variables el comportamiento de estos ante diferentes escenarios de fallas, dando como resultados mecanismo robustos y flexibles ante cambios en las condiciones del nodo sensor y el ambiente que lo rodea.

La inclusión de información contextual en la calidad de los mismos súpero las expectativas y hace robusto el mecanismo de control de calidad superando dos

barreras: la selección de parámetros de calidad y la adaptabilidad ante cambios externos; además, su modelado es de vital importancia en búsqueda de interoperabilidad, aquí nos ayudan los estándares de la OGC en la definición de características claves y relevantes para definir los metadatos de control de calidad.

La arquitectura propuesta refleja la capa de control de calidad de los datos y la forma como interactúa con las demás capas del SAD considerando el contexto de variables agroclimatológicas para agricultura de precisión. Esto fue posibles gracias a situarse en el modelo físico de un SAD, para sobreponerlo con un modelo contextual de sensado y enriquecerlo con estándares abiertos e interoperables como el SWE.

La capa de control de calidad y sus diferentes niveles muestran un modelo de datos contextual compuesto por valores agroclimatológicos capturados por los sensores, información proveniente de otros niveles e información contextual almacenada como metadatos. Además, los mecanismos de control de calidad seleccionados tienen el conocimiento del contexto de sensado y sus salidas son datos controlados en calidad, que a partir de una serie de marcas de calidad que siguen las normas de la OMM, son transmitidas a un centro de datos donde se toman las decisiones del negocio.

La división por niveles de la capa de control de calidad permite una especialización en el manejo de datos resultantes que conducen a una mejora continua en el control de calidad de todo el SAD.

La importancia de un aprendizaje en los sistemas de adquisición de datos, busca que las mediciones en los sensores se deban actualizar dependiendo de la variabilidad de los parámetros ambientales que son objeto de estudio, además se genera un ciclo de mejora continua en el proceso de control de calidad soportado en información contextual local y validada con información contextual global.

La validación de la arquitectura en un ambiente real permitió evaluar el mecanismo de control de calidad y ofrecer una herramienta portable en ambientes Arduino para control de calidad de datos como la temperatura, que es un parámetro valioso en muchas aplicaciones de agricultura de precisión.

6.2 Trabajos futuros

El modelado de información contextual es una parte muy importante en el proceso de control de calidad, pero en nuestro caso no se implementó totalmente ya que estaba por fuera de nuestro alcance, es muy importante que este modelado ofrezca interoperabilidad ya que las WSN son heterogéneas en comunicaciones, modelos de datos y dispositivos hardware en el nodo sensor.

El manejo del estándar SWE para redes limitadas en recursos, se está generando una iniciativa con este estándar para el Internet de las cosas IoT, se llama *SensorThing* que busca extender el uso de estándares de la OGC que buscan interoperabilidad para ser vistos desde Internet, aquí se hace uso del estándar O&M para generar observaciones y del servicio de observación del sensor SOS para enviarlos a la nube ya sea en formato *JSON* con tecnologías *RESTFul*.

Otro aspecto importante es la implementación de aprendizaje por refuerzo que ayuda a subsanar los problemas de adaptación de parámetros ante cambios en las condiciones del nodo sensor y del sitio de observación, la selección de este mecanismo de aprendizaje es porque se realiza localmente y no demanda muchos recursos del nodo sensor.

La exploración de otras técnicas de clasificación como SVM y redes bayesianas se pueden complementar con las usadas, para que el mecanismo logre mayor robustez y eficiencia en cuanto a consumo de energía en el procesamiento y la comunicación, con alto grado de detección de anomalías.

Bibliografía

Akyildiz, I.F. et al., 2002. A survey on sensor networks P. Stavroulakis & M. Stamp, eds. *IEEE Communications Magazine*, 40(8), pp.102–114. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1024422.

Asada, G. et al., 1998. Wireless integrated network sensors: Low power systems on a chip. *Proceedings of the 24th European SolidState Circuits Conference*, 43(5), pp.9–16. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1470957.

Atzberger, C., 2013. Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs. *Remote Sensing*, 5(2), pp.949–981. Available at: http://www.mdpi.com/2072-4292/5/2/949/ [Accessed July 14, 2014].

Ballari, D., Wachowicz, M. & Callejo, M.A.M., 2009. Metadata behind the interoperability of wireless sensor networks. *Sensors*, 9(5), pp.3635–3651.

Borghi, S., Favaron, M. & Frustaci, G., 2011. Surface meteorological monitoring network at mesoscale β, γ and microscale α quality, reliability and representativeness requirements. In *2011 IEEE Workshop on Environmental Energy and Structural Monitoring Systems*. IEEE, pp. 1–6. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6067049 [Accessed May 10, 2014].

Botts, M. & Robin, A., 2007. OpenGIS ® Sensor Model Language (SensorML) Implementation Specification. *Design*, p.180.

Bult, K. et al., 1996. Low power systems for wireless microsensors. *Proceedings of 1996 International Symposium on Low Power Electronics and Design*, pp.17–21. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=542724.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(September), pp.1–58. Available at: http://portal.acm.org/citation.cfm?id=1541882\nhttp://dl.acm.org/citation.cfm?id=1541882\nhttp://portal.acm.org/citation.cfm?doid=1541880.1541882.

Christin, D. et al., 2011. *A survey on privacy in mobile participatory sensing applications*, Available at: http://linkinghub.elsevier.com/retrieve/pii/S0164121211001701 [Accessed November 28, 2013].

Cid, P. del et al., 2011. Expressing and Configuring Quality of Data in Multi-purpose Wireless Sensor Networks. *Sensor Systems and …*, (i). Available at: http://link.springer.com/chapter/10.1007/978-3-642-23583-2_7 [Accessed February 6, 2014].

Cox, S., 2011. Observations and measurements-XML implementation. *OGC document*, pp.1–76. Available at: http://portal.opengeospatial.org/files/?artifact_id=44722\nhttps://publications.csiro.au/rpr/pub?list=BRO&pid=csiro:EP115858&sb=RECENT&n=11&rpp=50&page=89&tr=5033&dr=all&dc4.browseYear=2011.

Dereszynski, E.W. & Dietterich, T.G., 2011. Spatiotemporal Models for Data-Anomaly Detection in Dynamic Environmental Monitoring Campaigns. *ACM Trans. Sen. Netw.*, 8(1), p.3:1–3:36. Available at: http://doi.acm.org/10.1145/1993042.1993045\nhttp://dl.acm.org/ft_gateway.cfm?id=1993045&type=pdf.

Dey, A.K., 2001. Understanding and Using Context. *Personal and Ubiquitous Computing*, 5(1), pp.4–7. Available at: http://link.springer.com/10.1007/s007790170019 [Accessed November 12, 2013].

Dutta, R. et al., 2014. Development of an intelligent environmental knowledge system for sustainable agricultural decision support. *Environmental Modelling & Software*, 52, pp.264–272. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1364815213002351 [Accessed March 24, 2014].

Elsts, A. et al., 2012. SADmote: A robust and cost-effective device for environmental monitoring. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 225–237. Available at: http://link.springer.com/chapter/10.1007%2F978-3-642-28293-5_19#.

Van Essen, D.C. et al., 2012. The Human Connectome Project: a data acquisition perspective. *NeuroImage*, 62(4), pp.2222–31. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3606888&tool=pmcentrez&rendertype=abstract [Accessed March 30, 2014].

Estévez, J., Gavilán, P. & Giráldez, J. V., 2011. Guidelines on validation procedures for meteorological data from automatic weather stations. *Journal of Hydrology*, 402(1–2), pp.144–154.

Familiar, M.S. et al., 2012. Building service-oriented Smart Infrastructures over Wireless Ad Hoc Sensor Networks: A middleware perspective. *Computer Networks*, 56(4), pp.1303–1328.

Farruggia, A., 2011. *A probabilistic approach to anomaly detection for Wireless Sensor Networks Abstract*.

Fiebrich, C. a. et al., 2010. Quality Assurance Procedures for Mesoscale Meteorological Data. *Journal of Atmospheric and Oceanic Technology*, 27(10), pp.1565–1582. Available at: http://journals.ametsoc.org/doi/abs/10.1175/2010JTECHA1433.1 [Accessed February 6, 2014].

Fiebrich, C. a. & Crawford, K.C., 2001. The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bulletin of the American Meteorological Society*, 82(10), pp.2173–2187.

Gasparin, C.P., 2009. *an Exploratory Study Into the Acceptance of on Farm Automated Traceability Systems*.

Gwilliams, C., Preece, A. & Hardisty, A., 2012. Local and global knowledge to improve the quality of sensed data. *International Journal of …*, 2(2), pp.164–180. Available at: http://sdiwc.net/digital-library/local-and-global-knowledge-to-improve-the-quality-of-sensed-data [Accessed March 27, 2014].

Hamada, A. & Yatagai, A., 2011. An automated quality control method for daily rain-gauge data. *Global Environ. Res*, pp.183–192. Available at: http://www.airies.or.jp/attach.php/6a6f75726e616c5f31352d32656e67/save/0/0/15_2-12.pdf [Accessed February 6, 2014].

Hill, D.J. & Minsker, B.S., 2010. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, 25(9), pp.1014–1022. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1364815209002321 [Accessed August 26, 2014].

Huang, G. et al., 2014. Research on Data Quality of E&P Database Base on Metadata-Driven Data Quality Assessment Architecture. *Applied Mechanics and Materials*, 530–531, pp.813–817. Available at: http://www.scientific.net/AMM.530-531.813.

Hubbard, K., You, J. & Shulski, M., 2012. Toward a Better Quality Control of Weather Data. , pp.3–30. Available at: http://www.intechopen.com/books/practical-concepts-of-quality-control/toward-a-better-quality-control-of-weather-data [Accessed February 6, 2014].

IBRL, 2004. Intel Berkely Reseach Lab Dataset. Available at: http://db.csail.mit.edu/labdata/labdata.html [Accessed June 20, 2015].

Janakiram, D. et al., 2006. Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks. *2006 1st International Conference on Communication Systems Software & Middleware*, pp.1–6. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1665221.

Ji, S., Beyah, R. & Li, Y., 2011. Continuous data collection capacity of wireless sensor

networks under physical interference model. *Mobile Adhoc and Sensor Systems (MASS) …*. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6076620 [Accessed February 6, 2014].

Journée, M. & Bertrand, C., 2011. Quality control of solar radiation data within the RMIB solar measurements network. *Solar Energy*, 85(1), pp.72–86. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0038092X10003270 [Accessed February 6, 2014].

Jurdak, R. et al., 2011. Wireless Sensor Network Anomalies: Diagnosis and Detection Strategies. *Intelligent Systems Reference Library*, 10, pp.309–325.

Konieczny, M., 2012. ENRICHING WSN ENVIRONMENT. *Computer Science*, 13(4), pp.101–114. Available at: https://doaj.org/article/9ff73402d8124d6cad2b89f13861e4a9.

Kulkarni, R. V., Forster, A. & Venayagamoorthy, G.K., 2011. Computational intelligence in wireless sensor networks: A survey. *Communications Surveys & Tutorials, IEEE*, 13(1), pp.68–96. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5473889\nhttp://ieeexplore.ieee.org/ielx5/9739/5709955/05473889.pdf?tp=&arnumber=5473889&isnumber=5709955\nhttp://ieeexplore.ieee.org/xpl/login.jsp?reload=true&tp=&arnumber=5473889&url=http://ieeexplore.

LAU, C., Jarvis, A. & Ramírez, J., 2011. Agricultura Colombiana: Adaptación al Cambio Climático. *Centro Internacional de Agricultura Tropical (CIAT). 4p. …*, 1, p.4. Available at: http://ciat.cgiar.org/wp-content/uploads/2012/12/politica_sintesis1_colombia_cambio_climatico.pdf [Accessed August 27, 2014].

Lee, W.S. et al., 2010. Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture*, 74(1), pp.2–33. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168169910001493 [Accessed March 20, 2014].

Lemmens, R. et al., 2011. *New generation Sensor Web Enablement.*, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3231615&tool=pmcentrez&rendertype=abstract [Accessed November 21, 2013].

Libelium, 2013. Waspmote Plug & Sense. *Waspmote Technical Guide*, p.170. Available at: http://www.libelium.com/es/ [Accessed October 1, 2013].

Liu, Y., 2014. *A Cross-Layer Design for Sensor-based Ambient Intelligence Systems*. Auckland University of Technology.

Mason, S.J.K. et al., 2014. A centralized tool for managing, archiving, and serving point-in-time data in ecological research laboratories. *Environmental Modelling and Software*, 51, pp.59–69.

Meratnia, N. & Havinga, P., 2010. Outlier Detection Techniques for Wireless Sensor

Networks: A Survey. *IEEE Communications Surveys & Tutorials*, 12(2), pp.159–170. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5451757.

Ministerio, E., Agricultura, D. & Rural, D., 2006. La Agricultura de Precisión una oportunidad para la competitividad y la sostenibilidad de los sistemas productivos.

Muller, C.L. et al., 2013. Toward a Standardized Metadata Protocol for Urban Meteorological Networks. *Bulletin of the American Meteorological Society*, 94(8), pp.1161–1185. Available at: http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00096.1 [Accessed February 5, 2014].

Na, A. & Priest, M., 2007. Sensor Observation Service A. Na & M. Priest, eds. *English*, OGC 06-009(OGC 06-009r6), pp.1–104. Available at: http://portal.opengeospatial.org/files/?artifact_id=26667&passcode=pcq7e0gzzeat5n7erwhr.

National Oceanic and Atmospheric Administration, 1998. Automated Surface Observing System (ASOS) User's Guide. *Program Manager*, (March).

Ngai, E.C.-H. & Gunningberg, P., 2014. Quality-of-information-aware data collection for mobile sensor networks. *Pervasive and Mobile Computing*, 11, pp.203–215. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1574119213000941 [Accessed March 29, 2014].

Ni, K. et al., 2009. Sensor network data fault types. *ACM Transactions on Sensor Networks*, 5(3), pp.1–29.

Oliveira, L.M. & Rodrigues, J.J., 2011. Wireless Sensor Networks: a Survey on Environmental Monitoring. *Journal of Communications*, 6(2), pp.143–151. Available at: http://ojs.academypublisher.com/index.php/jcm/article/view/4233 [Accessed November 18, 2013].

Organizacion Meteorológica Mundial - OMM, 2011. *Guía de prácticas climatológicas Edición de 2011 OMM N° 100*, Ginebra Suiza. Available at: http://www.wmo.int/pages/prog/wcp/ccl/guide/documents/wmo_100_es.pdf.

Palpanas, T., 2013. Real-time data analytics in sensor networks. *Managing and Mining Sensor Data*, pp.1–29. Available at: http://link.springer.com/chapter/10.1007/978-1-4614-6309-2_7 [Accessed February 6, 2014].

Peets, S. et al., 2012. Methods and procedures for automatic collection and management of data acquired from on-the-go sensors with application to on-the-go soil sensors. *Computers and Electronics in Agriculture*, 81, pp.104–112. Available at: http://dx.doi.org/10.1016/j.compag.2011.11.011.

Percivall, G., Reed, C. & Davidson, J., 2007. Open Geospatial Consortium Inc . OGC White Paper OGC ® Sensor Web Enablement : Overview And High Level Architecture . *2007 IEEE Autotestcon*, 4540(December), pp.1–14. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4374243.

Rassam, M. a., Zainal, A. & Maarof, M.A., 2013. Advancements of data anomaly

detection research in Wireless Sensor Networks: A survey and open issues. *Sensors (Switzerland)*, 13(8), pp.10087–10122.

Ravichandran, J. & Arulappan, A.I., 2013. Data validation algorithm for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 2013(iv), p.11. Available at: http://www.hindawi.com/journals/ijdsn/2013/634278/abs/ [Accessed February 6, 2014].

Resch, B., Mittlboeck, M. & Lippautz, M., 2010. Pervasive monitoring--an intelligent sensor pod approach for standardised measurement infrastructures. *Sensors (Basel, Switzerland)*, 10(12), pp.11440–67. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3231049&tool=pmcentrez&rendertype=abstract [Accessed December 9, 2013].

Riopaila Castilla S.A., 2013. Agricultura de Precisión. Available at: http://www.riopailacastilla.com/index.php?option=com_content&view=article&id=33&Itemid=38 [Accessed August 26, 2014].

Schmidt, A., 2002. Ubiquitous Computing–Computing in Context. *Lancaster University, UK*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.1488&rep=rep1&type=pdf.

Suthaharan, S. et al., 2010. Labelled data collection for anomaly detection in wireless sensor networks. *2010 Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp.269–274.

Thessler, S. et al., 2011. Geosensors to support crop production: current applications and user requirements. *Sensors (Basel, Switzerland)*, 11(7), pp.6656–84. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3231660&tool=pmcentrez&rendertype=abstract [Accessed November 13, 2013].

Tseng, Y.-C., Wu, F.-J. & Lai, W.-T., 2013. Opportunistic data collection for disconnected wireless sensor networks by mobile mules. *Ad Hoc Networks*, 11(3), pp.1150–1164. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1570870513000036 [Accessed March 30, 2014].

U.S. Climate Data, U.S. Climate Data. Available at: http://www.usclimatedata.com/climate/berkeley/california/united-states/usca0087/2004/3 [Accessed June 30, 2015].

Vejen, F. et al., 2002. Quality Control of Meteorological Observations. , pp.1–109.

Wang, Q. & Balasingham, I., 2010. Wireless sensor networks-an introduction. *Wireless Sensor Networks: Application- …*, (187857). Available at: http://www.researchgate.net/publication/221909889_Wireless_Sensor_Networks_-_An_Introduction/file/79e4150e8748c1052c.pdf [Accessed March 4, 2013].

Xie, M. et al., 2011. Anomaly detection in wireless sensor networks: A survey. *Journal*

of Network and Computer Applications, 34(4), pp.1302–1325. Available at: http://dx.doi.org/10.1016/j.jnca.2011.03.004.

Yau, K.-L.A., Komisarczuk, P. & Teal, P.D., 2012. Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues. *Journal of Network and Computer Applications*, 35(1), pp.253–267. Available at: http://dx.doi.org/10.1016/j.jnca.2011.08.007.