

Fortalecimiento de las capacidades del Instituto de Investigaciones Marinas y Costeras - INVEMAR- mediante la implementaci  n de la ciencia de datos en investigaci  n marina y costera

C  digo: 80740-003-2020

Anexo 2

Manual de Reglas de validaci  n y consistencia series de datos   reas coralinas Informe t  cnico de avance

**Instituto de Investigaciones Marinas y Costeras
"Jos   Benito Vives de Andr  s" - INVEMAR
Santa Marta D.T.C.H., Marzo de 2021**



TABLA DE CONTENIDO

TABLA DE CONTENIDO.....	2
FIGURAS.....	3
TABLAS.....	3
OBJETIVO.....	4
ALCANCE	4
1. DESCRIPCIÓN	4
2. ROLES Y USUARIOS.....	5
3. APLICACIÓN DE BANDERAS DE CALIDAD.....	5
4. RECOLECCIÓN DE LA INFORMACIÓN.....	6
PLANTILLA: recolección de datos	7
5. VALIDACIÓN: INGRESO PLANTILLAS DE DATOS	10
6. DEPURACIÓN DE LOS CONJUNTOS DE DATOS (BASE DE DATOS SISTEMA MONITOREO DE CORALES DE COLOMBIA)	11
6.1. Condiciones previas	11
6.2. Proceso de depuración genérico.....	12
6.2.1. Fase de detección de errores	12
6.1.2. Fase de tratamiento	16
6.1.3 Fase de verificación	17
6.1.4 Fase de tratamiento	17
6.1.6 Fase de verificación	17
Generación de Datos para ICTAC y reporte final de inconsistencias.	17
7. VALIDACION DEL ICTAC	17
Descripción de la muestra (ICTAC)	18
8. REGLAS DE VALIDACIÓN Y CONSISTENCIA DE DATOS	18
9. RIESGOS	29
9. PARTICIPANTES	¡Error! Marcador no definido.



FIGURAS

Figura 1. Procesamiento de la información asociada a la operación estadística ICTAC	6
---	---

TABLAS

Tabla 1. Campos de la plantilla de datos SAMP_Estructura_Arrecifal.xlsm	7
Tabla 2. Campos de la plantilla de datos SAMP_Abundancia_peces_Arrecifal.xlsm	8
Tabla 3. Validaciones de los campos presentes en la VM_1480_88 estructura donde se encuentra el dataset para la estimación de la operación estadística ICTAC	18
Tabla 4. Validaciones de los campos presentes en la VM_1480_89 estructura donde se encuentra el dataset para la estimación de la operación estadística ICTAC	23
Tabla 5. Riesgos y controles asociados a los procesos de validación y consistencia	29





OBJETIVO

Establecer e implementar las reglas de validación y consistencia para los atributos y variables de la base de datos de Monitoreo de corales, que permitan detectarlas inconsistencias y errores que se pueden presentar en el ingreso y almacenamiento de información que soporta la operación estadística (OE) condición tendencia áreas coralinas, buscando garantizar la confiabilidad en la generación de datos estadísticos.

ALCANCE

Describe los procedimientos de validación y depuración a ejecutar en los microdatos para garantizar su calidad.

1. DESCRIPCIÓN

El indicador de la operación estadística condiciones tendencia áreas coralinas se calcula a partir de datos colectados en campo (*in situ*) y resultados de ensayos de laboratorio que se almacenan en la base de datos del Corales, que está diseñado y ejecutado mediante procesos programados PSL/SQL integrando diversos atributos que caracterizan la información. La implementación de las reglas de validación en el sistema permitirá optimizar los procedimientos de ingreso de datos minimizando la probabilidad de errores en las plantillas de carga masiva. Facilitará la revisión y depuración de datos, con la inclusión de criterios para validar si los atributos corresponden con los datos ingresados, de esta manera se minimizan las inconsistencias en las tablas materializadas y la posterior consulta de datos que son el insumo de los productos estadísticos.



2. ROLES Y USUARIOS

El proceso de validación y consistencia de datos dentro de corales, en el cual se soporta la operación estadística de condiciones tendencia áreas coralinas, estará bajo la responsabilidad del administrador temático encargado jefe de línea organización y dinámica de ecosistemas y por los jefes del laboratorio de servicios de información LABSIS del INVEMAR, quien contará con el apoyo de los siguientes usuarios:

- El usuario administrador temático, involucrado verifica que todos los datos hayan sido ingresados y depurados, dará apoyo a todo el proceso de implementación de las reglas de validación y consistencia.
- Investigadores estarán involucrados, principalmente en alimentar el sistema de información. Apoyarán la implementación de las reglas de validación y consistencia, ofrecerán el conocimiento temático y apoyo necesario para resolver dudas.
- Profesional LABSIS encargado de dar apoyo, proporcionar herramientas y ejecutar algunas de las acciones orientadas a la implementación de las reglas de validación y consistencia.
- Líder operación estadística: estará involucrado en el proceso y ofrecerá el conocimiento temático y apoyo necesario para resolver dudas, establecer criterios de expertos, y tomar decisiones en cuanto al proceso de validación y depuración de datos cuando se requiera.

3. APLICACIÓN DE BANDERAS DE CALIDAD

De acuerdo con estándares internacionales (IOC, 2013) a los datos se les asigna una bandera de calidad que consta de un rango entre 1 y 9, entre más bajo el número, se considera mejor la calidad del dato, estos valores son:

- 1= Bueno;
- 2= No evaluado o en proceso de evaluación;
- 3= Dudoso, dato que puede ser usado, pero con precaución;
- 4= Malo;
- 9= No reportado.

Solo los datos con bandera de calidad 1 o 3 confirmados pueden ser parte de los datos base para el cálculo del ICTac.

4. RECOLECCIÓN DE LA INFORMACIÓN

El proceso de recolección de información de la operación estadística ICTAC incluye actividades de monitoreo en campo, las cuales se describen en el protocolo indicador condición tendencia áreas coralinas (Rodríguez-Rincón, A. M., et al. 2014). El control de calidad de las actividades de recolección de datos y muestras en campo, incluyendo el aseguramiento de la calidad, son tareas asumidas por el Programa de Biodiversidad y Ecosistemas Marinos BEM, el cumplimiento de los protocolos es evaluado por el SGC de INVEMAR.

En la recolección de datos e información se utilizan diferentes formatos y registros que documentan el proceso desde la programación de la salida de campo, la captura de datos ambientales en campo, recolección de muestras, los cuales son diligenciados por el personal técnico idóneo asignado previamente en cada parte del proceso descrito en la Figura 1. Procesamiento de la información asociada a la operación estadística ICTAC:

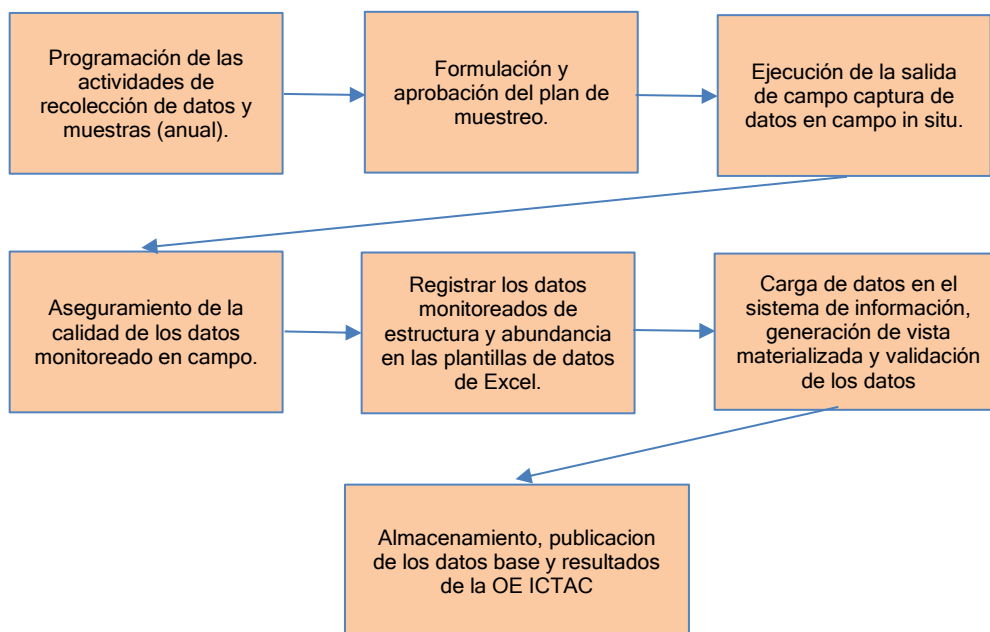


Figura 1. Procesamiento de la información asociada a la operación estadística ICTAC

PLANTILLA: recolección de datos

En las plantillas de Excel se registran los datos de monitoreo de corales para posteriormente cargarlo al sistema de información de monitoreo de los arrecifes coralinos en Colombia. Estos consisten en dos libros de Excel de nombres:

- "SAMP_Estructura_Arrecifal.xlsm": donde se registran los datos de estructura.
- "SAMP_Abundancia_Peces.xlsm": donde se registran los datos de abundancia de peces.

Las plantillas de datos de Excel cuenta con una pestaña llamada DATOS, en donde se registra la información del monitoreo y otra pestaña llamada DIGITADOR es donde se registra los datos relacionados con el proyecto, la fecha de creación de la plantilla, la persona que verifica y ejecuta la carga de datos al sistema también se describe las recomendaciones generales de diligenciamiento y de proceso.

Esta plantilla cuenta con macros que se conectan por medio de un servicio REST (representational state transfer), para la carga de datos al sistema de información de Información de Monitoreo del SIAM.

A continuación, se describe en detalle los campos que conforma la plantilla de datos de estructura:

Tabla 1. Campos de la plantilla de datos SAMP_Estructura_Arrecifal.xlsm

Pestaña	Campo	Descripción
DIGITADOR	Código	Código de la plantilla asignado en el sistema de monitoreo de los arrecifes coralinos en Colombia
	Nombre	Nombre de quien digita la información
	Entidad	Sigla de la entidad que genera el reporte. Por ejemplo: INVEMAR
	Plantilla	Versión de plantilla
	Fecha:	Fecha en la que ingresan los datos
	Código proyecto	Código generado automáticamente al seleccionar el nombre del proyecto
	Nombre	Nombre del proyecto
	Nombre del archivo	Nombre del archivo
	DOI Metadato	Url del conjunto de datos

	Fuente Plantilla	Url del repositorio donde se puede descargar la plantilla
DATOS	Codmuestreo	Indicar el valor numérico único del muestreo código asignado automáticamente
	Fecha	Fecha en la que ingresan las muestras
	Hora	Hora a la que se recolecta la muestra.
	PREF.	Prefijo de la estación asignado en el sistema de monitoreo de los arrecifes coralinos en Colombia
	COD.	Código de la estación asignado en el sistema de monitoreo de los arrecifes coralinos en Colombia
	ESTACION	Nombre de la estación generado automáticamente
	CODBDD	
	Datos tomados por	Nombre del personal responsable de ejecutar el muestreo, debe coincidir con el personal reportado.
	Entidad:	Sigla de la entidad que genera el reporte. Por ejemplo: INVEMAR
	Nivel	Profundidad de donde se realiza el monitoreo
	Transecto	Número del transecto de donde se realiza el monitoreo
	Ubicación	Concatenación del nivel y transecto generado automáticamente.
	PUNTO/VALOR	Valor del punto intercepto donde se captura muestra generado automáticamente.
	Cat_vida	Categoría de vida de la muestra monitoreada
	Descripción	Campo que se diligencia automáticamente al digitar un código valido de cat_vida
	Especie	Selección de la sigla de la especie
	Nombre de especie	Nombre científico de la especie y se diligencia automáticamente al digitar un código valido de especie

A continuación, se describe en detalle los campos que conforma la plantilla de datos de abundancia de peces.

Tabla 2. Campos de la plantilla de datos SAMP Abundancia_peces_Arrecifal.xlsm

Pestaña	Campo	Descripción
DIGITADOR	Código	Codificación en la base de datos para el nombre quien digita la información.
	Nombre	Nombre de quien digita la información
	Entidad	Sigla de la entidad que participan en el monitoreo. Por ejemplo: INVEMAR
	Plantilla	Versión de plantilla
	Fecha:	Fecha en la que digitan los datos
	Código proyecto	Código generado automáticamente al seleccionar el nombre del proyecto
	Nombre	Nombre del proyecto
	Nombre del archivo	Nombre del archivo
	DOI Metadato	Url donde se describe el conjunto de datos que se está documentando
	Fuente Plantilla	Url del repositorio donde se puede descargar la plantilla
DATOS	Codmuestreo	Indicar el valor numérico único del muestreo código asignado automáticamente
	FECHA	Fecha en la que digitan los datos
	HORA	Hora a la que se recolecta la muestra.
	DATE	Concatenación de la fecha y la hora este registro se genera automáticamente.
	PREF.	Prefijo de la estación asignado en el sistema de monitoreo de los arrecifes coralinos en Colombia
	COD.	Código que hacer referencia a profundidad o si es expuesto o protegido
	ESTACION	Nombre de la estación generado automáticamente
	CODBDD	Código de la estación en la base de datos generado automáticamente
	MAR	El mar donde se monitorea
	DATOS TOMADOS POR	Nombre del personal responsable de ejecutar el muestreo, debe coincidir con el personal reportado.
	ENTIDAD	Sigla de la entidad que genera el reporte. Por ejemplo: INVEMAR
	NIVEL	Profundidad de donde se realiza el monitoreo
	UBICACIÓN	Concatenación del nivel y transecto generado automáticamente.

	OBSERVACION TRANSECTO	Observaciones adicionales a la muestra
	BANDA	Número de banda donde se monitorea
	CDG ESPECIE	Selección de la sigla de la especie
	N. científico	Nombre científico de la especie y se diligencia automáticamente al digitar un código valido de especie
	CANTIDAD	Número de especie muestreada
	LONGITUD	Rango de longitud a la que pertenece la especie muestreada.
	LONGITUD(cm)	Campo que se diligencia automáticamente al digitar un rango de LONGITUD
	FAMILIA	Nombre de la familia que se diligencia automáticamente al digitar un código valido de especie

5. VALIDACIÓN: INGRESO PLANTILLAS DE DATOS

El proceso de validación y la digitación de datos inicia con el diligenciamiento de los libros de recolección de datos

- "SAMP_Estructura_Arrecifal.xlsm"
- "SAMP_Abundancia_Peces.xlsm"

disponible en http://cinto.invemar.org.co/download/Plantillas_ARGOS_Plus/. Es necesario tener en cuenta que, las celdas sombreadas en gris se encuentran programadas para realizar la validación y si se registran errores le indicará el número de errores en la tabla de resultado llamada REGISTROS GENERADOS en la hoja DIGITADOR.

Luego se procede a copiar los datos en la pestaña Datos, se completan los datos de la pestaña Digitador, y se procede a ejecutar las macros Generar Tablas y Cargar Datos para estructurar y llevar los datos al sistema. Si alguna de las dos macros previas arroja errores se verifican de acuerdo a los mensajes que envíe el sistema y si es necesario se recurre al apoyo del profesional del LABSIS que acompaña el proceso. En particular las macros dejan algunos datos de trazabilidad relacionadas con el número de registros y de presentarse errores los códigos de errores enviados por el sistema para cada uno de los procesos. Esos datos se encuentran en las columnas A, B, C filas 19 y subsiguientes de la pestaña digitador.

La carga de datos implica algunas validaciones adicionales que provienen de

reglas de integridad aplicadas por el administrador de base de datos, ellas aplican para:

- ID Proyecto,
- ID Metodología
- ID Investigadores
- ID Estaciones o puntos de muestreo

6. DEPURACIÓN DE LOS CONJUNTOS DE DATOS (MODULO SISTEMA MONITOREO DE CORALES DE COLOMBIA)

6.1. Condiciones previas

Para que el proceso de depuración sea más eficiente, conviene antes de ejecutarlo verificar que se han ingresado todas las plantillas de datos y asegurarse que, de acuerdo al procedimiento de backup para el sistema, se tenga una copia del esquema de la base de datos de manera que sea posible revertir modificaciones o eliminaciones no deseadas de una parte considerable de los registros

Todo dato en la plantilla de dato Excel para cargar tendrá asignado una bandera de calidad de 2 (en verificación), la cual cambiará de acuerdo al resultado del proceso de validación.

Los procedimientos a ejecutar se desarrollaron usando PL/SQL sobre la base de datos ORACLE, R como paquete estadístico de programación y Excel como herramienta auxiliar para el reporte y documentación del proceso de depuración de datos.

El proceso de depuración de datos se realiza a partir de los datos primarios almacenados en el sistema de monitoreo de corales de Colombia, los cuales se encuentran organizados en la vista materializada:

- VM_AGM_1480_88 para estructura
- VM_AGM_1480_89 para abundancia

del esquema DATOS DE CAMPO del gestor de base de datos ORACLE, instancia del Sistema de Información Ambiental Marina — SIAM. Posterior a la depuración de dicha vista materializada, se genera la vista VM_AGM_1480_88DC de la cual se extraen los datos que sirven de insumo para el cálculo del ICTAC.

6.2. Proceso de depuración genérico

6.2.1. Fase de detección de errores

Una vez terminado el proceso de carga de datos, el investigador ejecuta el procedimiento para generar la vista materializada correspondiente, con la opción de hacer actualizaciones incrementales (solo los registros ingresados desde un año en particular). De este primer ciclo a la vista Materializada se agrega una columna denominada "ERR_CREACION", en la que en formato texto, se listan errores relacionados con:

- Campos obligatorios que se encontraron nulos o campos con caracteres para su tipo de datos no admisibles que son convertidos a nulos
- Campos con tipo de dato equivocado, por ejemplo, números sistema decimal con caracteres impropios o fechas incorrectas
- Códigos no declarados en los listados de dominio correspondientes

A cada uno de estos errores que el investigador puede filtrar en la vista lógica:

- VM_DC_8800_CHECKS_BASIC0 para estructura.
- VM_DC_8900_CHECKS_BASIC0 para abundancia.

Estos corresponden al nivel de calidad 4, se reportan con el prefijo ERR- seguido del código de la variable y un texto explicativo. El investigador aplica la corrección de los mismos.

El investigador dispone además de un conjunto de vistas lógicas, que deberá ir revisando en forma ascendente hasta depurar y documentar los errores, cada una de ellas tiene los mismos atributos que la VM_AGM_1480_88DC, más una columna denominada ERROR_DC que describe el error en específico, estas son:

➤ PARA LOS DATOS ESTRUCTURA

- VM_DC_8801_TRANSECTO: Registros con error de número de puntos por transecto reportado no mayor a lo que especifica el protocolo de indicador Samp, transectos válidos de acuerdo a las estaciones definidas.
- VM_DC_8802_CUMP_ESP_CAT_VIDA: Registros que reportan datos con categorías de vida equivocadas para la especie que se registró en la muestra. A estos registros les corresponde la bandera

de calidad 4.

- VM_DC_8803_ FECHA_MONITO: Registros con error en la fecha del monitoreo o es mayor a la fecha calendario actual o es menor a la fecha de las primeras colectas de datos de monitoreo de calidad. A estos registros les corresponde la bandera de calidad 4.
- VM_DC_8803_ DUPLICADOS: Registros que reportan datos para la misma estación, variable, fecha y categoría de vida, especie.
- VM_DC_8804_LIMITES_APROX: Registros que reportan datos que de acuerdo a criterio de experto son extremos o improbables, o alertas para datos que de acuerdo a criterio de experto son probables pero atípicos.

➤ PARA LOS DATOS DE ABUNDANCIA

- VM_DC_8901_CUMP_BANDA: Registros con error de numero de bandas monitoreada mayor a lo que especifica el protocolo de indicador Samp.
- VM_DC_8902_CUMP_ESP: Registros que reportan datos con especies equivocadas, repetidas y cumplimiento para el mar que se está monitoreando. A estos registros les corresponde la bandera de calidad 4.
- VM_DC_8903_ FECHA_MONITO: Registros con error en la fecha del monitoreo o es mayor a la fecha calendario actual o es menor a la fecha de las primeras colectas de datos de monitoreo de calidad. A estos registros les corresponde la bandera de calidad 4.
- VM_DC_8903_ DUPLICADOS: Registros que reportan datos duplicados para la misma estación.
- VM_DC_8904_LIMITES_APROX: Registros que reportan datos que de acuerdo a criterio de experto son extremos o improbables, o alertas para datos que de acuerdo a criterio de experto son probables pero atípicos.

El investigador o el encargado del proceso documenta el proceso de identificación de errores y correcciones, donde se lleva el registro y control los cambios ejecutados para la corrección de los errores y las verificaciones frente a las alertas. Las correcciones deben aplicarse preferiblemente siguiendo la secuencia de las vistas lógicas presentada arriba (columna de cada vista llamada ERROR_DC).

Validación de completitud

El proceso de validación devuelve a partir de una vista lógica a la base de datos un formato Excel de control en el que se listan organizados por fecha-mes, estación, el número de variables reportadas, y una cadena de texto, en la que se organizan por grupo y luego alfabéticamente las variables con datos. Este proceso se realiza para garantizar el cumplimiento del estándar de

COMPLETITUD.

Eliminación espacios y caracteres no imprimibles del texto

El procedimiento verifica que no existan caracteres "invisibles" que, alteran el significado de los datos blancos o caracteres que provienen de copiar y pegar objetos digitales de diversas fuentes. Para realizar esta tarea se utiliza un algoritmo de validación de campos, basado en expresiones regulares (secuencias de caracteres que conforman patrones de búsqueda). Este conjunto de expresiones se utiliza comúnmente en el campo de la informática para el diseño de algoritmos de análisis léxico y permite validar en la base de datos que un campo contenga únicamente los tipos de caracteres permitidos, como, por ejemplo, letras o números, excluyendo así caracteres simbólicos o extraños.

Validación de atributos con base a reglas

Entre los criterios de validación asignados para la VM_AGM_1480_88DC se encuentran la No Nulidad de campos, relación de llaves primarias (ej. Para X sector solo son válidos Y códigos de estación), rangos numéricos permitidos por variable, rangos de fecha permitidos, formato horario, tipo de dato (numérico o carácter), longitud permitida de caracteres y valores permitidos por variable según las escalas, estándares u opciones acogidas por el instituto (es necesario aclarar que la información contenida dentro de la base de datos es de dos tipos, cualitativa y cuantitativa), por lo cual muchos de los criterios de validación y depuración van a ser totalmente distintos.

Se validan las reglas del Diccionario de Datos para:

- Evaluar la No Nulidad de campos para los atributos de la vista materializada VM_AGM_1480_88DC referenciados en el DICCIONARIO DE DATOS.
- Evaluar que el valor de las variables categóricas tenga relación con los valores descritos en las listas de referentes AG_LOV.
- Verificar la relación de INTEGRIDAD para las entidades, los proyectos, los usuarios y estaciones, para los cuales existen objetos definidos en la base de

datos.

- Los atributos con formato de fecha, deben cumplir con los criterios básicos relacionados con la secuencia de tiempo en la que se ejecuta cada actividad. Para el caso de las fechas almacenadas como texto se verifica que efectivamente se trate

de fechas, el formato estándar es AAAA/MM/DD (AÑO cuatro dígitos, MES dos dígitos, DÍA dos dígitos) si el dato lleva hora será en el formato de 24 horas.

- Verificar las reglas de CONSISTENCIA entre atributos como la correlación de ID (X campo depende del valor ingresado en Y campo), que X método tenga unidad de medida.
- Verificar rangos numéricos o de caracteres permitidos por atributos (ej. mes solo números del 1 al 12), campos únicamente numéricos, las relaciones existentes entre unidades de medida
- Verificar límite de caracteres para algunos atributos.

En el caso de las variables cualitativas, existen algunos casos en donde éstas se representan a través de una secuencia fija de números que llega a ser considerada como un carácter numérico, por lo cual, para estos casos en particular, el algoritmo solo debe verificar que la cadena de caracteres esté compuesta por caracteres numéricos.

Las variables cualitativas presentan generalmente los formatos de carácter, texto y fecha, aclarando que el formato de tipo carácter estará condicionado por variable.

Para finalizar, se diseñó un algoritmo que notifica al administrador de la base de datos sobre el número de registros que se han revisado, el número de errores que se han encontrado y los tipos de estos errores (ej. Se han revisado 1000 registros de los cuales 50 presentan campos nulos, 35 presentan errores de fecha, 20 tienen categorización equivocada y 15 tienen errores de inconsistencias). El algoritmo realiza un comparativo entre los datos y el dominio válido para el respectivo atributo. En el caso de que el elemento no se encuentre dentro del dominio de opción para esa variable, el algoritmo genera un reporte de notificación al administrador de la base de datos, en donde se muestran todas las inconsistencias encontradas. Este proceso tiene como finalidad mantener la INTEGRIDAD, VALIDEZ y CONSISTENCIA de la base de datos.

El investigador descargara en formato Excel cada una de las vistas (VM_DC..) descrita, por medio de la herramienta sqldeveloper para posteriormente revisar los errores, luego en el libro descargado en la hoja de los datos se crea una columna de nombre SOLUCION_DC y especifica las correcciones que haya lugar y las aclaraciones, con base a eso se decide qué hacer con el registro.

Luego se debe de guardar el libro de Excel en el Microsoft SharePoint de ODI con las correcciones cambiando el nombre del libro agregando la versión (Ej. Nombre_archivo_corregida_v1.1), este archivo se guarda en la carpeta de VALIDACION ICTAC/2021, con el fin de llevar la trazabilidad del proceso validación de datos

6.1.2. Fase de tratamiento

El profesional IABSIS encargado, tomando como insumo las vistas lógicas y reportes entregados por los procedimientos aplicados en la fase de detección, recopila las plantillas de datos fuente que sean necesarias e inicia la actividad de chequear uno por uno los errores, determinando las acciones a tomar, las cuales pueden ser:

- Borrar registros duplicados
- Modificar contenidos de registros que aparentan estar duplicados pero que contrastados con datos fuente demuestran errores en los datos fuente.
- Corregir los errores puntuales en los registros
- Borrar los datos de una plantilla totalmente, hacer las correcciones sobre los datos fuente y llevarlos nuevamente al sistema.
- Reportar las inconsistencias relacionadas a las unidades reportadas, haciendo las anotaciones pertinentes. En el caso de que se desee que el sistema haga la corrección, se desarrollará un proceso para esto. Si por el contrario se desean hacer manualmente bajo criterio de expertos, este cambio debe ser ingresado en la base de datos. Por último, se documentará el cambio de unidades que se realizó.
- Consultar para registros con valores atípicos y ajustar su bandera de calidad a 3 del tipo ALRT o cuatro.

Para cada acción se debe coordinar con los usuarios involucrados y se registra en el reporte de la inconsistencia y la acción tomada.

6.1.3 Fase de verificación

En esta fase se actualiza la vista materializada VM_AGM_1480_88DC y se vuelve a ejecutar el proceso de depuración genérico. Por último, se ejecuta el procedimiento que asigna banderas de calidad a cada uno de los registros. Los únicos registros que cambiarán de bandera de calidad por procedimiento son los que tengan bandera de calidad 2 que se actualizarán a uno.

6.1.4 Fase de tratamiento

A partir del reporte generado en la etapa de diagnóstico, profesional IABSIS encargado de los datos revisa nuevamente contra los datos fuentes, para descartar errores en las cifras. Si estos persisten, se tiene la opción de evaluar las mediciones con el personal investigador de apoyo, de tal forma que, estos determinen si el registro presenta un dato verdadero extremo, verdadero normal (es decir, que la expectativa previa era incorrecta) o si realmente se presenta un error de medición.

Como resultado final se asignan las banderas de calidad pertinentes para cada caso, clasificando al registro bueno o con valor dudoso.

6.1.6 Fase de verificación

El administrador de datos actualiza la vista materializada y repite, si es necesario, el proceso de análisis estadístico con las ecuaciones ya programadas en la base de datos.

Generación de Datos para ICTAC y reporte final de inconsistencias.

Una vez hecha la estandarización sobre los archivos limpios, se actualiza la vista materializada de la base de datos VM_1480_88 para estructura y VM:1480_89 para abundancia que da origen a las vistas con los indicadores de la operación estadística ICTAC.

El profesional IABSIS encargado llevará el registro de control los cambios ejecutados sobre los registros para los procesos de validación. Este reporte final contendrá el reporte inicial de los elementos que no pasaron los criterios de validación de campos, el reporte de registros con datos atípicos y un reporte de actualización, en el que se definen aquellos registros que han sido actualizados con su fecha de actualización. Adicionalmente, también se generará un reporte con aquellos datos que no fueron actualizados.

A partir de la V_SAMP_INDICADOR ICTAC, se ejecutan las transformaciones estadísticas necesarias para constituir las nuevas unidades estadísticas que no son provistas de forma explícita en la recolección, pero que se necesitan para obtener los resultados requeridos. Las unidades estadísticas y el ICTAC son sometidos a un proceso de depuración que comprende las etapas que se describen a continuación.

Descripción de la muestra (ICTAC)

El líder de la operación estadística y el equipo temático realiza una revisión de los datos disponibles para la estimación del ICTAC, y los valida aplicando los siguientes criterios:

- Comprueba la disponibilidad de datos para los puntos de muestreo definidos para el reporte
- Verifica que todos los datos tengan la estimación de la transformación con la aplicación de las curvas ajustadas definidas para cada variable
- Valida los resultados atípicos, comprobando el dato registrado

8. REGLAS DE VALIDACIÓN Y CONSISTENCIA DE DATOS

En la Tabla 8.1 se describen las reglas de validación y consistencia que deben cumplir los datos para el cálculo de la operación estadística del ICTAC, donde la columna campo corresponde al nombre de cada uno de los atributos que se encuentran en la base de datos y la validación indica todas las reglas que debe ser tenidas en cuenta para validar cada dato.

Las tablas referentes señaladas a continuación se encuentran los dataset para la estimación de la operación estadística ICTAC

Descripción de los atributos de: Tabla 3. Validaciones de los campos presentes en la **VM_1480_88** estructura donde se encuentra

el dataset para la estimación de la operación estadística ICTAC

Tabla 3. Validaciones de los campos presentes en la VM_1480_88 estructura donde se encuentra el dataset para la estimación de la operación estadística ICTAC

Campo	Regla de Validación	Observaciones	Tabla Referente
ID_MUESTREO	1.Campo requerido, NO DEBE SER NULO 2.Campo con valor único	Número Consecutivo Número	N/A



	3.Debe ser de tipo de datos numérico	Consecutivo para la campaña de muestreo en el año	
ID_MUESTREOTX	1.Campo requerido, NO DEBE SER NULO 2.Campo con valor único 3.Debe ser de tipo de datos alfanumérico	Número Consecutivo para la campaña de muestreo en el año	N/A
ID_ESTACION	1.Campo requerido, NO DEBE SER NULO 2.El código de la estación debe estar previamente definido en el listado de dominios válidos para las estaciones 3.Debe ser de tipo de datos numérico	Código de la estación	Referencia en el esquema Geográficos en la tabla CESTACIONES_RAZIZ
NOM_ESTACION	1. Campo requerido, NO DEBE SER NULO 2. El código de la estación debe estar previamente definido en el listado de dominios válidos para las estaciones 3. El nombre de la estación debe corresponder al ID_ESTACION ingresado	Nombre de la estación monitoreada	Referencia en el esquema Geográficos en la tabla CESTACIONES_RAZIZ
ID_PROYECTO	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para este campo	Proyecto al que pertenecen los datos ingresados	Referencia en el esquema Geográficos en la tabla CLST_PROYECTOS
ID_METODOLOGIA	1.Campo requerido, NO DEBE SER NULO 2.campo numerico	Metodología aplicada en el monitoreo	
ID_TEMATICAS	1.Campo requerido, NO DEBE SER NULO	Temática a la que se muestrea	Referencia en el esquema Datos de campo en la tabla AGM_TEMATICA
FECHA	1.Campo requerido, NO DEBE SER NULO		N/A

	2. Fechas en formato dd/mm/yyyy 3. Las fechas deben estar en el rango entre el 01/01/1990 y la fecha actual del sistema	Fecha del muestreo	
FECHA_ANO	1. Campo requerido, NO DEBE SER NULO 2. Debe tener una longitud de 4 Debe ser numérico entero	Año en el que se realizó el monitoreo Se extrae del campo FECHA_REAL	N/A
FECHA_MES	1. Campo requerido, NO DEBE SER NULO 2. Debe tener una longitud de 7 caracteres de tipo Alfanumérico	Mes en el que se realizó el monitoreo Se extrae del campo FECHA_REAL	N/A
PARTICIPANTES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombres de los participantes quien captura el dato en campo y quien digitaliza los datos	Referencia en el esquema Datos de campo en la tabla AGD_ATORIAS
ENTIDADES	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para éste campo	Nombre de la entidad que ejecuta el monitoreo	Referencia en el esquema GEOGRAFICOS
NOTAS_GENERALES	1. Campo no requerido 2. Campo alfanumérico	Observaciones para el muestreo	N/A
TEMPORADA	1. Campo requerido, NO DEBE SER NULO 2. Debe tener una longitud de 7 caracteres de tipo Alfanumérico	Código para la temporada estacional relacionada con la toma de la muestra	N/A
URL_METADATO	1. caracteres de tipo Alfanumérico	URI persistente para el metadato del conjunto de datos	N/A
VERSION_PLANTILLA	1. caracteres de tipo Alfanumérico	Versión para la plantilla en la que se	N/A

		organizaron los datos fuente	
ARCHIVO_FUENTE	1.caracteres de tipo Alfanumérico	Nombre archivo del que proviene el conjunto de datos	N/A
DES_ESTACION	1.caracteres de tipo Alfanumérico	Descripción de la estación monitoreada	N/A
COD_ESTACION	1. Campo requerido, NO DEBE SER NULO 2.EL código de la estación debe pertenecer al sector seleccionado 3.El código debe estar definido en el listado de estaciones	Código de la estación monitoreada	Referencia en el esquema GEOGRAFICOS
AREA	1. Campo requerido, NO DEBE SER NULO 2.Valor numérico 3.El código debe estar definido en el listado de areas	Código del Área monitoreada	Referencia en el esquema GEOGRAFICOS
AREA_DES	1. Campo requerido, NO DEBE SER NULO 2.caracteres de tipo Alfanumérico 3.El código debe estar definido en el listado de areas	Nombre del área monitoreada	Referencia en el esquema GEOGRAFICOS
LOCALIDAD	1. Campo requerido, NO DEBE SER NULO 2.Valor numérico 3.El código debe estar definido en el listado de localidades	ARCHIVO_FUENTE	Referencia en el esquema GEOGRAFICOS
LOCALIDAD_DES	1. Campo requerido, NO DEBE SER NULO 2.caracteres de tipo Alfanumérico 3.El código debe estar definido en el listado de localidades	Descripción de la localidad donde se monitorea	Referencia en el esquema GEOGRAFICOS
ID_MUESTRA	1. Campo requerido, NO DEBE SER NULO	Código del muestreo del monitoreo, debe ser	N/A

	2. Solo dígitos 0-9	numérico	
NOTAS	2.caracteres de tipo Alfanumérico	Observaciones para la muestra	
ES_REPLICA	1. Campo requerido, NO DEBE SER NULO 2. Debe ser un número entero mayor que 0	Hace referencia al único dato que se registró de la estación al momento del muestreo 1, 2, 3	N/A
QUALITY_FLAG	1. Campo no requerido 2. Campo numerico	una bandera de calidad que consta de dos niveles, en un rango entre 1 y 9	N/A
ENTIDAD	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para este campo.	Código de la entidad que realiza el monitoreo	Referencia en el esquema GEOGRAFICOS
ENTIDAD_DES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombre de la entidad que realiza el monitoreo	Referencia en el esquema GEOGRAFICOS
UBICACION_ESPECIFICA	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico conformado por dos caracteres	Transecto y profundidad del muestreo	N/A
VALOR_ESLABON	1. es un campo requerido, NO NULO 2. Debe ser un valor numérico 3. NO negativo 4. NO Puede tener decimales	Punto intercepto de la captura de la muestra monitoreada	N/A
CAT_VIDA	1.Campo requerido, NO DEBE SER NULO 2.Debe estar definido previamente en el listado de dominios para este campo.	Siglas de la categoría de vida	Referencia en el esquema Datos de campo en la tabla AG_ATRIBUTOS_ADICIONALES
CAT_VIDA_DES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombre de la categoría de vida	Referencia en el esquema Datos de campo en la tabla AG_ATRIBUTOS_ADICIONALES

ESPECIE	1. acepta valores NULO 2. Debe estar definido previamente en el listado de dominios para este campo.	Código asignado para la especie evaluada	Referencia en el esquema Datos de campo en la tabla AG_ESPECIES
ESPECIE_DES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombre asignado para la especie evaluada	Referencia en el esquema Datos de campo en la tabla AG_ESPECIES
ERR_CREACION	1. Campo no requerido 2. Campo alfanumérico	Error de creación de la vista materializada informa los errores	N/A
FACTUALIZACION	1.Campo requerido, NO DEBE SER NULO 2. Fechas en formato dd/mm/yyyy 3. Las fechas deben estar en el rango entre el 01/01/1990 y la fecha actual del sistema	Fecha de la última actualización	N/A

La descripción de las reglas de validación para los datos de estructura se listan en la tabla 4.

Tabla 4. Validaciones de los campos presentes en la VM_1480_89 estructura donde se encuentra el dataset para la estimación de la operación estadística ICTAC

Campo	Regla de Validación	Observaciones	Tabla Referente
ID_MUESTREO	1.Campo requerido, NO DEBE SER NULO 2.Campo con valor único 3.Debe ser de tipo de datos numérico	Número Consecutivo Número Consecutivo para la campaña de muestreo en el año	N/A
ID_MUESTREOTX	1.Campo requerido, NO DEBE SER NULO 2.Campo con valor único 3.Debe ser de tipo de datos alfanumérico	Número Consecutivo para la campaña de muestreo en el año	N/A

ID_ESTACION	1.Campo requerido, NO DEBE SER NULO 2.El código de la estación debe estar previamente definido en el listado de dominios válidos para las estaciones 3.Debe ser de tipo de datos numérico	Código de la estación	Referencia en elesquema Geográficos en la tabla CESTACIONES_R AIZ
NOM_ESTACION	1. Campo requerido, NO DEBE SER NULO 2. El código de la estación debe estar previamente definido en el listado de dominios válidos para las estaciones 3. El nombre de la estación debe corresponder al ID_ESTACION ingresado	Nombre de la estación monitoreada	Referencia en elesquema Geográficos en la tabla CESTACIONES_R AIZ
ID_PROYECTO	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para este campo		
ID_PROYECTO	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para este campo	Proyecto al que pertenecen los datos ingresados	Referencia en elesquema Geográficos en la tabla CLST_PROYECTO S
ID_METODOLOGIA	1.Campo requerido, NO DEBE SER NULO 2.campo numerico	Metodología aplicada en el monitoreo	
ID_TEMATICAS	1.Campo requerido, NO DEBE SER NULO	Temática a la que se muestrea	Referencia en elesquema Datos de campo en la tabla AGM_TEMATIC A
FECHA	1.Campo requerido, NO DEBE SER NULO 2. Fechas en formato dd/mm/yyyy	Fecha del muestreo	N/A

	3. Las fechas deben estar en el rango entre el 01/01/1990 y la fecha actual del sistema		
FECHA_ANO	1. Campo requerido, NO DEBE SER NULO 2. Debe tener una longitud de 4 Debe ser numérico entero	Año en el que se realizó el monitoreo Se extrae del campo FECHA_REAL	N/A
FECHA_MES	1. Campo requerido, NO DEBE SER NULO 2. Debe tener una longitud de 7 caracteres de tipo Alfanumérico	Mes en el que se realizó el monitoreo Se extrae del campo FECHA_REAL	N/A
PARTICIPANTES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombres de los participantes quien captura el dato en campo y quien digitaliza los datos	Referencia en el esquema Datos de campo en la tabla AGD_ATORIAS
ENTIDADES	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para éste campo	Nombre de la entidad que ejecuta el monitoreo	Referencia en el esquema GEOGRAFICOS
NOTAS_GENERALES	1. Campo no requerido 2. Campo alfanumérico	Observaciones para el muestreo	N/A
CLASE_SUSTRATO		Código clase medio físico del que se toma la muestra	
TEMPORADA	1. Campo requerido, NO DEBE SER NULO 2. Debe tener una longitud de 7 caracteres de tipo Alfanumérico	Código para la temporada estacional relacionada con la toma de la muestra	N/A
VERSION_PLANTILLA	1. caracteres de tipo Alfanumérico	Versión para la plantilla en la que	N/A



		se organizaron los datos fuente	
ARCHIVO_FUENTE	1.caracteres de tipo Alfanumérico	Nombre archivo del que proviene el conjunto de datos	N/A
AREA	1. Campo requerido, NO DEBE SER NULO 2.Valor numérico 3.El código debe estar definido en el listado de areas	Código del Área monitoreada	Referencia en elesquema GEOGRAFICOS
DES_ESTACION	1.caracteres de tipo Alfanumérico	Descripción de la estación monitoreada	N/A
COD_ESTACION	1. Campo requerido, NO DEBE SER NULO 2.EL código de la estación debe pertenecer al sector seleccionado 3.El código debe estar definido en el listado de estaciones	Código de la estación monitoreada	Referencia en elesquema GEOGRAFICOS
AREA	1. Campo requerido, NO DEBE SER NULO 2.Valor numérico 3.El código debe estar definido en el listado de areas	Código del Área monitoreada	Referencia en elesquema GEOGRAFICOS
AREA_DES	1. Campo requerido, NO DEBE SER NULO 2.caracteres de tipo Alfanumérico 3.El código debe estar definido en el listado de areas	Nombre del área monitoreada	Referencia en elesquema GEOGRAFICOS
LOCALIDAD	1. Campo requerido, NO DEBE SER NULO 2.Valor numérico 3.El código debe estar definido en el listado de localidades	ARCHIVO_FUENTE	Referencia en elesquema GEOGRAFICOS
LOCALIDAD_DES	1. Campo requerido, NO DEBE SER NULO	Descripción de la localidad donde se monitorea	Referencia en elesquema GEOGRAFICOS

	2.caracteres de tipo Alfanumérico 3.El código debe estar definido en el listado de localidades		
ID_MUESTRA	1. Campo requerido, NO DEBE SER NULO 2. Solo dígitos 0-9	Código del muestreo del monitoreo, debe ser numérico	N/A
NOTAS	1.caracteres de tipo Alfanumérico	Observaciones para la muestra	
ES_REPLICA	1. Campo requerido, NO DEBE SER NULO 2. Debe ser un número entero mayor que 0	Hace referencia al único dato que se registró de la estación al momento del muestreo 1, 2, 3	N/A
QUALITY_FLAG	1. Campo no requerido 2. Campo numérico	una bandera de calidad que consta de dos niveles, en un rango entre 1 y 9	N/A
REGION	1. tipo Alfanumérico de máximo un carácter	Región a la que se monitoreo	N/A
ENTIDAD	1. Campo requerido, NO DEBE SER NULO 2. Debe estar definido previamente en el listado de dominios para este campo.	Código de la entidad que realiza el monitoreo	Referencia en eleschema GEOGRAFICOS
ENTIDAD_DES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombre de la entidad que realiza el monitoreo	Referencia en eleschema GEOGRAFICOS
UBICACION_ESPECIFICA	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico conformado por dos caracteres	Transecto y profundidad del muestreo	N/A
CENSO	1. es un campo requerido, NO NULO 2. Debe ser un valor numérico 3. NO negativo 4. NO Puede tener decimales	Banda donde se captura las muestra	N/A



	5.valor no mayo de 10		
ESPECIE	1. acepta valores NULO 2. Debe estar definido previamente en el listado de dominios para este campo.	Código asignado para la especie evaluada	Referencia en elesquema Datos de campo en la tabla AG_ESPECIES
ESPECIE_DES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Nombre asignado para la especie evaluada	Referencia en elesquema Datos de campo en la tabla AG_ESPECIES
CANTIDAD	1. es un campo requerido, NO NULO 2. Debe ser un valor numérico 3. NO negativo 4. NO Puede tener decimales	Numero de ocurrencia de peces muestreada de la misma especie	N/A
LONGITUD	1. es un campo requerido, NO NULO 2. Debe ser un valor alfanumérico 3. valores entre la A - C 5.valor no mayo de 10	Rango de longitudes de una especie muestreada	N/A
LONGITUD_DES	1. Campo requerido, NO DEBE SER NULO 2. Campo alfanumérico	Descripción del rango de la longitud muestreada	N/A
URL_METADATO	1.caracteres de tipo Alfanumérico	URI persistente para el metadato del conjunto de datos	N/A
ERR_CREACION	1. Campo no requerido 2. Campo alfanumérico	Error de creación de la vista materializada informa los errores	N/A
FACTUALIZACION	1.Campo requerido, NO DEBE SER NULO 2. Fechas en formato dd/mm/yyyy 3. Las fechas deben estar en el rango entre el 01/01/1990 y la fecha actual del sistema	Fecha de la última actualización	N/A

9. RIESGOS

En La Tabla 5. Riesgos y controles asociados a los procesos de validación y consistencia se describen los riesgos y controles asociados a las validaciones y apartados descritos en el presente documento.

Tabla 5. Riesgos y controles asociados a los procesos de validación y consistencia

Riesgo	Control
Ingresar de información en la plantilladesactualizada	Incluir la versión en la plantilla y asociarla en el sistema (verificar en el sistema cuales datos están ingresados en determinada la versión)
Ingresar variables o valores para campos con dominios definidos no validos	Rutina de validación en la base de datos y controles dentro de la plantilla mediante macros programadas
Pérdida accidental de datos en la base de datos	Capacitación en manejo de la herramienta de administración de base de datos
Pérdida de las plantillas de datos estandarizados	Llevar un registro electrónico de las plantillas estandarizadas al sistema, donde se especifique el departamento, proyecto, variables, quien envía los datos

9. BIBLIOGRAFIA

IOC, Intergovernmental Oceanographic Commission "Ocean Data Standards Volume 3. Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data. Version 1." Report. UNESCO-IOC, 2013. <https://doi.org/10.25607/OBP-6>.

Rodríguez-Rincón, A. M., et al. 2014. Protocolo Indicador Condición Tendencia Áreas Coralinas ICTAC. INVEMAR, Serie de Publicaciones Generales N° 66, <https://doi.org/10.21239/V9R594>.

10. Autor

NOMBRE	CARGO	FECHA
JEISON A. DIAZ PALMERA	PROFESIONAL DE APOYO LABSIS	MARZO 18 DE 2121