

Control de errores en Datos Geográficos

Carlos López Vázquez
carlos.lopez@ieee.org

Enfoque

- Existen cursos de Control de Calidad
 - Típicamente para Ingeniería Industrial o Química
- Métodos concebidos para productos tangibles
- Apuntan a:
 - Cadenas de producción en serie
 - Medir atributos de la Calidad
 - Detectar problemas rápidamente...
 - Actuar eliminando la causa
- Con datos es un poco más difícil...

Con datos...

- Muchas veces no salen de una *cadena*
 - Ejemplos:
 - Mapa de una ciudad
 - Catastro urbano
 - Contraejemplos:
 - Datos meteorológicos
 - Series o colecciones cartográficas
 - Imágenes satelitales
- Lo normal es que ya *estén* recogidos o generados
 - No incidiremos en su creación ☺

Problemas...

- Problemas principales:
 - ¿Cómo saber si un dato/registro está bien?
 - ¿Nos afecta?
 - Si está mal: ¿cómo asignarle un sustituto?
- Problemas secundarios:
 - ¿Cómo medir la Exactitud?
 - ¿Cómo comunicar la Calidad?
 - Una vez lograda, ¿cómo mantener la Calidad en una producción en serie?
- Por cierto... ¿Qué es *exactamente* Calidad?

Anecdotario...

- Ilustra algunos ejemplos de la vida real
- Barbas del vecino arder...
- Son ejemplos geográficos; hay muchos más en otras áreas

El BMW último modelo...

Fuente: P. Fisher (1999)

- En Alemania, 1998 un BMW se precipitó en el río Havel en el embarcadero de un ferry
- El dueño seguía fielmente las instrucciones de un sistema GPS con mapas instalado a bordo
- El mapa decía "puente" y debió decir "ferry"



Embajada china en Belgrado

- Bombardeada por la OTAN en 199X por error
- Los mapas decían que era una instalación serbia
- Había sido... ¡antes!
- Tres muertos, roces, etc.

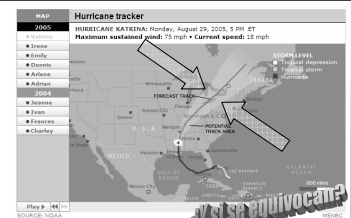


Derrame de petróleo

- Punta del Este, Uruguay en febrero de 1997
- Escollo de 24 mt. no figuraba en las cartas náuticas oficiales
- Un petrolero se lo llevó por delante



Huracán Katrina



Más casos marítimos

- 1980, Canadá. El *Sea Fever* pierde un marinero en una tormenta
- El pronóstico decía que la tormenta no llegaría donde llegó
- El error se debió a una boya meteorológica fuera de servicio
- El litigio por US\$ 3.2 millones fue ganado



Fuente: Richard O. Mason (1986)

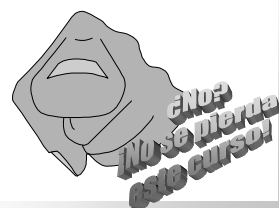
Hay casos más antiguos...

- El canal de Suez existió en la época faraónica
- Napoleón no lo reconstruyó porque se creía que $\Delta h \approx 6$ mts.
- Esa diferencia requería esclusas, etc.
- En realidad son 80 cm

La India podría haber caído en otras manos... y la historia hubiese sido diferente



¿Ud. quisiera estar en esta lista?



Temario

- Definiendo la Calidad → 4 h
- Comunicando la Calidad → 4 hs
- Midiendo la Exactitud → 6 hs
- Mejorando la Exactitud → 12 hs
- Manteniendo la Calidad → 4 hs

Total previsto: 30 horas

¿Por qué tanto?

Definiendo la Calidad (4 hs.)

- Hay que comenzar por el principio
- Terminología, jerga, etc.
- Quizá sea familiar al público (en parte...)
- Dar contexto de datos geográficos
 - Estándares recientes
 - Informatización *pervasiva*
 - Revolución en varias profesiones
- Lo que es de *alta calidad* para uno ...

Comunicando la Calidad (4 hs.)

- ¿Alguien ha escuchado de la INDE?
- ¿Alguien ha escuchado de *metadatos*?
- ¿Alguien ha visto algún metadato?
- Veremos qué es, cómo interpretarlo y qué buscar en relación a Calidad
- Revisaremos los estándares vigentes



Midiendo la Exactitud (6 hs.)

- ¿Qué métodos o estándares se aplican?
- Caso posicional:
 - Los que saben son uds.
 - Revisaremos algún estándar vigente
- Caso temático:
 - Hay métodos pero ...
 - Veremos criterios académicos
- Servirá para saber
 - Si tenemos que preocuparnos (*problem dependent!*)
 - Si efectivamente *mejoramos*...



Mejorando la Exactitud (12 hs.)

- Tema *muy* nuevo
- Ámbito académico (1996, 1998,...)
- Grupo pequeño
- Tema *demasiado* técnico
 - Journals estadísticos no lo publican mucho
 - Journals temáticos (p.ej. Meteorológicos) lo ignoran
- ¿A quién le importa?
 - Productor del dato
 - Usuario del dato
 - (sufrido usuario debí decir...)
- Constituye el núcleo *duro* del curso



Mejorando la Exactitud (cont.)

- Las preguntas básicas son:
 1. ¿Cómo detectar un dato malo?
 2. ¿Cómo corregirlo?
- La segunda es más trillada: interpolación, mínimos cuadrados, etc.
- La primera... tiene cola
 - ¿Existe un (único) *valor correcto*?
 - ¿Es accesible? ¿A qué precio?
- ¿Cómo hago?
- ¿Lo que dijo Juan, no será para peor?



Manteniendo la Calidad (4 hs.)

- ¿Cómo asegurar niveles estables de error?
- Importante para:
 - Compradores de datos a proveedores externos
 - Productores de datos
 - Académicos (*of course...*)
- Tema nuevo en Agrimensura: típico y *tradicional* de Industriales o Químicos
- Veremos estándares, métodos, etc.

Referencias y materiales

- Ariza, 2002
- López *et al.*, 1999
- CD con material vario
 - PPT con "transparencias"
 - PDF consultados
 - Informes de proyecto
 - Tesis
 - Páginas y sitios web integros
 - Quizá algo de software
 - Etc....

Módulo 1: Definiendo la Calidad

Carlos López Vázquez

carlos.lopez@ieee.org

Plan

- Definiendo la Calidad
- Exactitud vs. Precisión
- Componentes de la Calidad



Algunos problemas técnicos...

- Por ejemplo:
 - dar coordenadas a objetos
 - cómo crear representaciones digitales
 - cómo procesar esas representaciones
- Hay otros...
 - error, incertidumbre, escala, resolución...
- Complexivamente: *Calidad de datos*

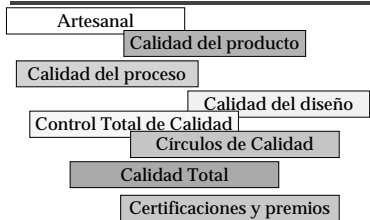
¿Qué es Calidad?

- RAE:

Propiedad o conjunto de propiedades inherentes a una cosa que permiten apreciarla como igual, mejor o peor que las restantes de su especie.
- Otra definición: ISO 9000

Grado en que un producto o servicio cumple con los requisitos especificados. No se refiere a atributos, generalmente implícitos, sino a la relación de orden.

Evolución histórica



Exactitud vs. Precisión

- Exactitud (Geodetic Glossary):

Es una cercanía de los resultados, cálculos o estimaciones a los valores verdaderos o valores que se acepta son verdaderos
- El idioma español...

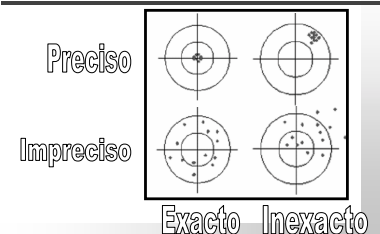
Precisión ≠ Exactitud
- Hay varias "Precisión"
 - Precisión (NCGIA Core):

Se define como el número de cifras decimales o de dígitos significativos de una medida
 - Precisión (Geodetic Glossary):

En estadística, una medida de la tendencia de un conjunto de números aleatorios a agruparse alrededor de un número determinado por el conjunto



Ilustrando Exactitud vs. Precisión



Más sobre Exactitud...

- A veces no hay "valor verdadero..."
- $E(x)$ no es lo mismo que $E(f(x))$...
 - Ej.: MDE vs. Pendientes
- Hay Exactitud *alta* y *baja*
- No son sólo medidas. Comparar:
 - Un valor exacto ($\alpha+\beta+\gamma=180$)
 - Una unidad convencional (metro patrón)
 - Un valor que se asume más exacto (monolito)

Precisión (NCGIA)

- Suele ser >>Exactitud
- Herencia informática...
 - Cifras significativas ≠ Cifras previstas
 - Es fácil confundirse
 - PC/ARCINFO → Single Precision
 - Algunos GIS → Sólo Integer
- Es un error relativo
- Más relacionada con Resolución
- Usada por Burroughs, 1986 y otros
- No es la definición más moderna, ni la más apropiada

Precisión (Geodetic Glossary)

- Tiene en cuenta la repetibilidad
 - +Precisión ↔ -Varianza
- Un aspecto clave:
 - Precisión depende sólo del juego de datos
 - En cambio la Exactitud no
- Para confundir más...
 - "Nivelación de alta precisión"
- Es la definición que usaremos

Los problemas son...

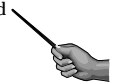
- Cómo medir la Exactitud
 - ¡Repetir las medidas no ayuda!
- Cómo se propaga la Exactitud
- Cómo no atribuir más Exactitud de la correcta
 - Es un problema del lado del *usuario*

Otro concepto: Resolución

- Mezclable con Precisión (NCGIA)
- Se distinguen tres casos:
 - Espacial
 - Temática
 - Temporal
- No abundaremos...

Plan

- ✓ Definiendo la Calidad
- ✓ Exactitud vs. Precisión
- Componentes de la Calidad



Componentes de la Calidad

- Recogida en varios estándares
 - CEN/TC 287
 - ISO/TC 211
 - USGS 1994
 - ISO 19113
- Hay muchos elementos comunes
 - Exactitud Posicional
 - Exactitud Temática
 - Consistencia lógica
 - Completitud
 - Linaje



Exactitud Posicional

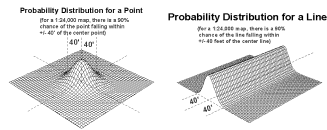
- Definida como la cercanía a la posición verdadera
- Usualmente involucra coordenadas
- Convención: 0.5 mm a la escala del plano
 - Ej.: $0.5 / (1/10000) = 0.5 \times 10^4 \text{ mm} = 50 \text{ m}$
 - En general: 0.5 / escala
 - ¿Escala? ¡Regla para el mundo en papel!
- Cifras significativas...
Ej.: 1:24.000; (x,y)=(123456.789,987654.321) mts.
Error=0.5/(1/24000)=12.0 mts → sólo valen algunos dígitos
(1.2345x10⁵, 9.8765x10³)

¿Y si no hay/hubo papel?

- Se debe declarar la "Exactitud"
- ¿Cómo?
 - La posición de los puntos difiere de la verdadera en menos de X mts.
 - Pregunta: ¿Cuáles puntos? ¿Todos?
 - USGS: En el 90% de los casos, la posición difiere de la verdadera en menos de X mts.
 - Aparece el percentil 90%
 - No asume una distribución a priori
 - BSI: La desviación estándar del error es menos de X mts.
- ¿Cómo interpretar la declaración?
- ¿Cómo verificarla?

Cómo interpretar...

■ Según USGS



Cómo verificarla...

■ Tres alternativas:

- Usar una fuente de mayor Exactitud
 - GPS
 - Mapa de mayor escala
 - Volver a los datos originales
- Usar evidencia interna
 - Tamaño de los defectos
- Calcular la Exactitud propagando...
 - 0.5 mm por aquí, 1 mm por allá...

■ Será desarrollado en detalle luego

Exactitud de Atributos

- También definida como *cercanía al valor verdadero*
- Atención: el tiempo ahora pesa más
 - Aparece la Vigencia
- ¿Cómo se expresa la Exactitud? Depende del tipo de dato:
 - Atributos continuos
 - Ej.: MDE, pluviometría, etc.
 - Parecido al caso posicional
 - Ej.: La elevación tiene un error menor a 1 mt
 - Podrían estar categorizados (imágenes RGB, etc.)
 - Atributos categóricos

Para Atributos Categóricos

■ Definir "cercanía" es diferente...

- Está definido el "#", pero no el "-"
- Podría estar definido el ">"

■ Hay otras diferencias

- ¿Existe un "verdadero valor"?
- ¿Cómo son las categorías?
 - ¿Apropiadas?
 - ¿Suficientemente detalladas?
 - ¿Bien definidas? ¿Nitidas?

■ También será analizado en detalle luego

Cómo verificarla...

■ Hay varias alternativas

- Construyendo una *matriz de confusión*
 - Presume una inspección binaria
 - Existen índices para caracterizarla
- Usando Métodos Difusos
 - La inspección ahora no es binaria
 - Aún bajo investigación
- Otros índices
 - Los hay, pero son algo simples...
- En cualquier caso, se hace Prueba de Hipótesis

■ Aún no está recogido en los estándares

Consistencia Lógica

■ ¿Es la BD consistente con sus definiciones?

- Geométricas:
 - Una etiqueta y sólo una por polígono
 - Todo polígono con etiqueta
 - Todo punto es de un polígono
 - Ningún arco que se cruza con otro sin cortarse
 - Tiene topología
 - Temáticas:
 - Los atributos están dentro del rango
- La Consistencia Lógica se mide en % de cumplimiento

Más sobre Consistencia...

■ ¿Fueron *todas* las capas obtenidas...?

- Misma escala, mismas hipótesis, mismos procesos...

■ ¿Fueron todas las capas editadas...?

- Mismos criterios (overshoots, tolerancias,...)

■ ¿Fueron todas las capas obtenidas simultáneamente?

- Hora, estación, año, etc. (lo que corresponda)

Complejidad

■ Difícilmente medible

■ No hay una definición formal

"...Dado el Modelo Cartográfico, la BD y la Realidad, los objetos que deberían estar efectivamente están..."

■ Calidad depende del Modelo Cartográfico

- Ej.: Incluir líneas de 30kV o más
- La BD está completa, aunque falten las de 6.3kV

■ Podría ser por Atributos

- El objeto está, pero incompleto
- Ej.: Falta la profundidad de un pozo

■ No hay métricas consensuadas

Linaje

■ Es un registro de historia de la BD

■ Responde a preguntas como:

- ¿Cómo fue digitalizada? ¿A partir de qué documentos?
 - ¿Cuándo fueron recogidos los datos base?
 - ¿Qué agencia u organización recolectó los datos?
 - ¿Qué pasos u etapas se utilizaron para procesar los datos originales?
 - ¿Con qué precisión fueron efectuados los cálculos (o qué error tenían los resultados numéricos)?
- El Linaje es usualmente un indicador útil de Exactitud
- No hay métricas consensuadas



Hay otros también relevantes...

- Vigencia
- Accesibilidad
- Pertinencia o Relevancia
- Son considerados en algunos estándares de metadatos
- Veremos de qué se trata

Vigencia

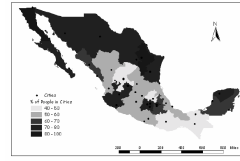
- ¿Están los datos actualizados?
 - Ej.: uso de tierra con fines de agricultura



Formato OK, Resolución OK, Completitud OK etc....

Otro ejemplo...

- Población urbana, México 1990



¿1965 vs. 1990? Puede ser mejor 1965...

Accesibilidad...

- Que los datos existan...



- Burocracia, formatos exóticos, etc.

Pertinencia o Relevancia

- Ciertos datos pueden no aportar nada...
- A veces se descubre; a veces no
 - Modelos importados
 - Condiciones locales
- Estadística tramposa; un ejemplo:
 - Idea: Búsqueda de individuos con mayor Respuesta a promociones
 - ≠ fdem con mayor propensión a pagar!!!

Módulo 1: Definiendo la Calidad

Carlos López Vázquez

carlos.lopez@ieee.org

Módulo 2 Comunicando la calidad: Metadatos

Carlos López Vázquez

carlos.lopez@ieee.org

¿Qué es un metadato?

- Describe el contenido, calidad, condición y otras características de los datos
- Usos principales de los metadatos:
 - organizar y preservar las inversiones en datos de una organización
 - proveer información a catálogos de datos y al ClearingHouse
 - proveer información para facilitar la transferencia de datos

¿Qué es un metadato?⁽²⁾

- Responde a las preguntas usuales:
 - ¿Qué?
 - ¿Cómo?
 - ¿Quién?
 - ¿Para qué?
 - ¿Dónde?
 - ...
- No es una redacción...

Tendencias en Metadatos

- Vincular los metadatos con archivos de datos, transferencia y sistemas de respaldo
- Deseo de usar tecnologías de SGBD y WEB para publicar metadatos
- Interés creciente en documentar recursos relevantes no digitales (p.ej. proyectos, servicios, expertise)



Estándares...

- También hay iniciativas para generar Metadatos de modelos matemáticos (software). Ver por ejemplo:
www.ncgia.ucsb.edu/projects/metadata/standard/standard.doc
Content Standard for Computational Models
Version 1.2
Metadata for Models Work Group
Alexandria Digital Earth Prototype Project
- Disponible en el CD

Introducción al estándar de Metadatos FGDC 1998

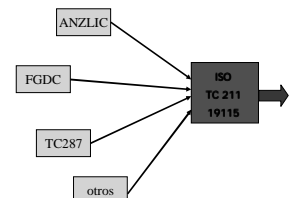
¿Porqué el FGDC 1998?

- Amplia experiencia existente
- Adopción internacional fuera de EEUU
 - Uruguay ☺
 - Australia-Nueva Zelanda
 - México, Canadá, etc.
- Bien documentado, con tutoriales
- Con software específico, gratuito
- La del artillero... (lo conozco en detalle...)

Antecedentes

- Desarrollo del estándar FGDC
 - Junio 8, 1994 Versión 1.0
 - Junio 19, 1998 Versión 2.0
- Desarrollo del estándar ISO
 - Noviembre 2000, Borrador Internacional del Estándar
 - Julio 2003, Versión final del Estándar Internacional
- Otros...

Convergencia



Objetivos (lo que hace)

- Da soporte a usos corrientes de los metadatos
 - inversiones internas - ClearingHouse - transferencia
- Desarrollado con la perspectiva de "¿Qué necesito saber sobre un juego de datos?"
 - disponible - acceso - transferencia - apropiado
- Provee un conjunto común de terminología y definiciones, así como información sobre valores a ser suministrados
- Identifica elementos obligatorios, obligatorios si corresponde y opcionales

Objetivos (lo que no hace)

- El estándar no especifica:
 - la forma de organizar la información dentro de una computadora o sistema
 - los medios para organizar la información para ser transferida
 - los medios a través de los que la información es transmitida, comunicada o presentada al usuario

Decisiones de Implementación

- ¿Qué es un "juego de datos"?
- ¿Cuándo es el mejor momento para generar los metadatos?
- ¿Para quién es todo esto?
 - Gerencia - catálogo - transferencia
 - Detalles, detalles y más detalles
 - El mundo no depende de Ud (¿o sí?)
- Datos preexistentes y el futuro

Aspecto de los metadatos

- Identificación
 - ¿Título? ¿Area cubierta? ¿Temas? ¿Actualización? ¿Restricciones?
- Calidad de Datos
 - ¿Nivel de error? ¿Compleitud? ¿Consistencia lógica? ¿Limpieza?
- Organización espacial de los datos
 - ¿Indirecta? ¿Vector? ¿Raster? ¿Tipo de elementos? ¿Número?
- Referencia espacial
 - ¿Proyección? ¿Sistema de cuadrícula? ¿Datum? ¿Sistema de coordenadas?
- Información de Entidad y Atributos
 - ¿Características? ¿Atributos? ¿Valores de los atributos?
- Distribución
 - ¿Distribuidor? ¿Formato? ¿Medio físico? ¿En línea? ¿Precio?
- Referencia de metadatos
 - ¿Actualización del metadato? ¿Redactor responsable?

Disección del estándar FGDC 1998

Elementos de la definición

- Secciones
- Elementos compuestos
- Elementos del dato

Secciones

- Constituyen los capítulos principales del estándar
- Están compuestos por:
 - Definición de la sección
 - Lista de elementos, definiciones, tipos y valores
 - Información sobre qué es obligatorio y repetible

Ejemplo de una sección

Información de Identificación - Información básica sobre el juego de datos

Tipo: compuesto
Nombre abreviado: idinfo

Información de Identificación =

Cita +
Descripción +
Periodo_Asociado_al_Contenido +
Status + ...

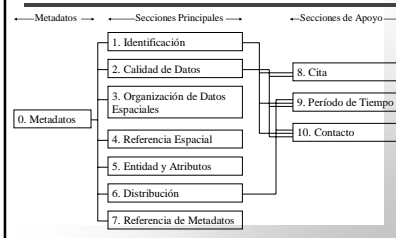
Cita - Información para ser usada para referirse al juego de datos

Tipo: compuesto
Nombre abreviado: citeinfo

Descripción - Una caracterización del juego de datos, incluyendo su uso previsto y limitaciones

Tipo: compuesto
Nombre abreviado: descript

Secciones del estándar



Elementos compuestos

- Es un grupo de elementos simples relacionados o de otros elementos compuestos
 - Todos los elementos compuestos están formados en última instancia por elementos simples
- **Formato:**
Nombre del elemento compuesto -- definición
Tipo: compuesto
- **Ejemplo:**
Descripción -- una caracterización del conjunto de datos, incluyendo su uso previsto y limitaciones
Tipo: compuesto

Elemento simple

- Una primitiva lógica de ítems de datos
 - Los elementos simples son los que uno rellena
- **Formato:**
Nombre del elemento simple -- definición
Tipo: (elegir entre "entero", "real", "texto", "fecha" u "hora")
Dominio: (lista los valores que pueden ser asignados)
- **Ejemplo:**
Resumen -- una breve descripción del conjunto de datos
Tipo: texto
Dominio: texto libre

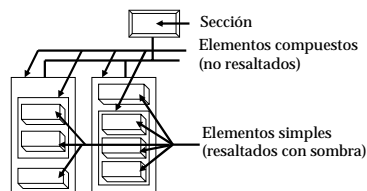
Valores válidos

- El *Dominio* de valores para un elemento simple:
 - puede ser sólo especificado por tipo. En estos casos, se agrega la palabra "libre" (texto libre, entero libre, etc.)
 - puede ser especificado por una lista, referencia a una lista o un rango
 - puede ser parcialmente especificado desde una lista, u opcionalmente ser libre

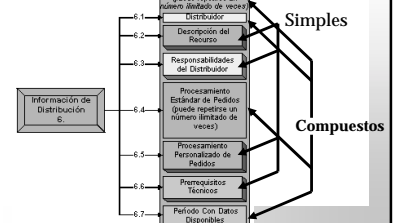
Casos especiales

- El estándar especifica el formato para cuatro casos
 - Fecha
 - Hora del día
 - Latitud y Longitud
 - Direcciones en Internet y nombres de archivos asociados

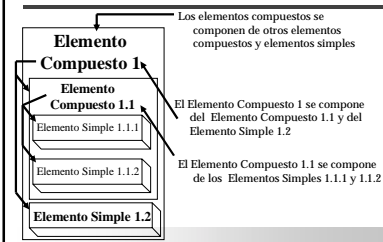
Representación gráfica de los elementos



Ejemplo: Distribución



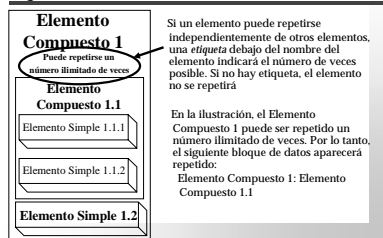
¿Cómo se agrupan?

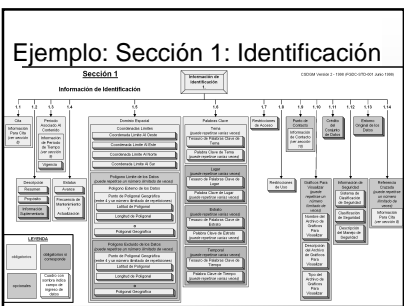
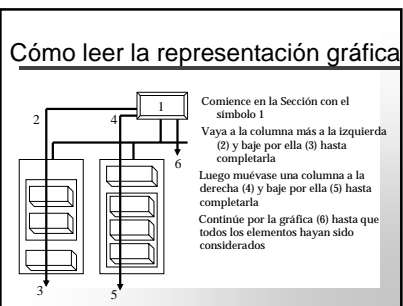
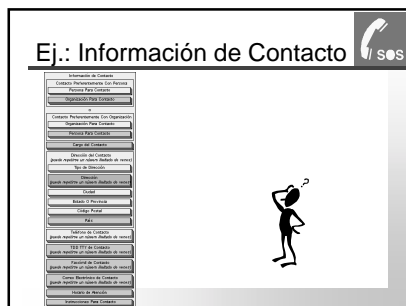
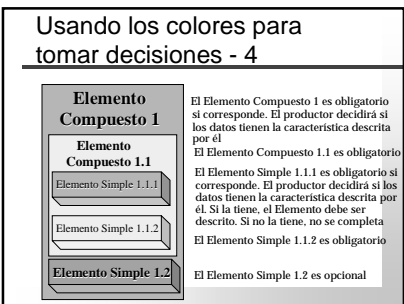
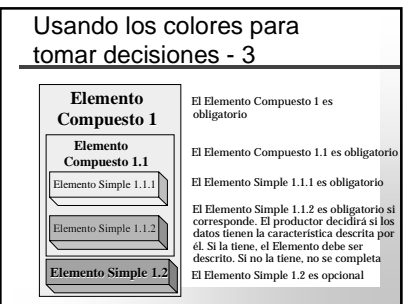
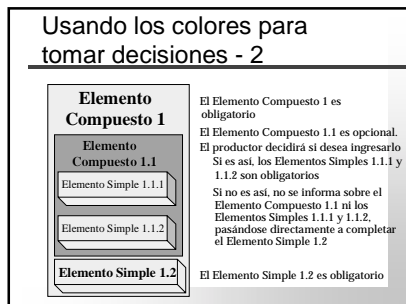
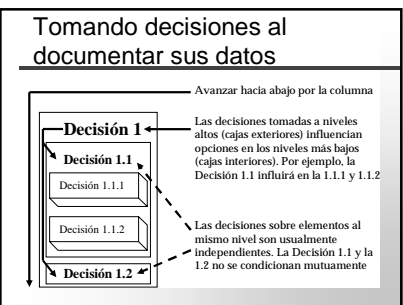


¿Qué es obligatorio? ¿Qué no?

Elemento Compuesto	Elemento Simple	Significado
		Obligatorio: debe ser completado
		Obligatorio si corresponde: debe ser completado si los datos exhiben la característica definida
		Opcional: suministrado a la discreción del productor de los datos

¿Qué puede repetirse? ¿Cuántas veces?

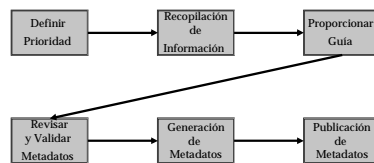




Muy lindo... ¿pero cómo lo hago?



¿Qué hacer y en qué orden?



Herramientas para Metadatos

Software para Metadatos

- Para la captura de METADATOS existen varias herramientas como son: MetaLite, CorpsMet95, Tkme, etc.
- MetaLite es un programa que sirve para crear y validar metadatos cubriendo *un conjunto mínimo* de elementos del FGDC.
- CorpsMet95 es otro programa que también permite crear y validar metadatos, cubriendo *todo el conjunto* de campos del FGDC.
- Tkme es un programa que permite crear metadatos, el cual también cubre *todo el conjunto* de campos del FGDC.

Validación de Metadatos: mp

- Es un analizador (Parser) de metadatos, que verifica la sintaxis y crea los archivos de salida.
- Los archivos serán nombrados por un prefijo (ejem.: sidf) con los sufijos .text, .sgml, y .html
 - sidf.text
 - sidf.sgml
 - sidf.html
- MP está disponible para Windows (MS-DOS), Linux y Unix.

Validación de Metadatos

Comando: mp.exe [-e efile] [-t tfile] [-s sfile] [-h hfile] arch_ent

arch_ent: es el nombre del archivo capturado.
 efile: archivo donde se graban los errores.
 archivos de salida: tfile archivo texto,
 sfile archivo sgml,
 hfile archivo html.

Ejemplo:

```

mp -e sidf.err sidf.txt
mp -t sidf.text -s sidf.sgml -h sidf.html sidf.txt (MetaLite)
mp -t sidf2.text -s sidf2.sgml -h sidf2.html sidf2.met (CorpsMet95)
  
```

Presentación de Metadatos

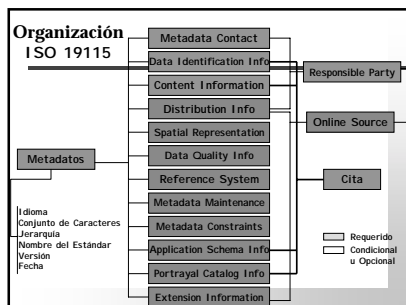
- Las salidas son de varios tipos
 - Versión WEB, para humanos
 - Versión WEB, para humanos expertos
 - Versión SGML, para indizar
 - Versión texto, para desarrollo/validación
- MP permite salidas más o menos bonitas
- Ejemplo:

**Aquí va un
ejemplo de
metadato terminado**

ISO 19115 La Próxima Generación

- Se conformó un consenso Internacional sobre metadatos estructurados dentro de un estándar de Geomática más amplio incluido en el Comité Técnico ISO 211 (TC 211)
- ISO 19155 recoge experiencias del FGDC, TC 287, ANZLIC y otros
- También especifica guías para el contenido (vocabulario y estructuras), como el FGDC





Metadatos *Full ISO 19115*

- Incluyen elementos de catálogos y estructuras
- Tienen previsto más detalles que el FGDC en particular para información raster y de imágenes
- En su mayor parte elementos y estructuras condicionales
- Luce familiar para los habituados al FGDC

Núcleo del ISO 19115

- Diseñado para ser usado en catalogar datos para su descubrimiento
- Incluye aproximadamente 50 campos
- Cubre los "quién", "qué", "cuándo", "dónde", "porqué", y "cómo" mejor que el núcleo del FGDC
- Puede no ser adecuado para otros requerimientos más detallados

Comparación FGDC ↔ ISO

- El nivel de detalle de metadatos del FGDC e ISO es comparable
- Los campos obligatorios en el FGDC y el Núcleo ISO son similares, aunque ISO espera más detalle
- FGDC tiene previsto construir herramientas de migración de FGDC V2 a ISO
- *ISO 19115 será conocido como FGDC V3*
Wait and see...

Módulo 2 Comunicando la calidad: Metadatos

Carlos López Vázquez

carlos.lopez@ieee.org

Módulo 3: Midiendo la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org

Plan

- Error, Exactitud y Precisión
- Definiendo estándares
- Informando sobre la Calidad
- Midiendo y Verificando
- Análisis de Sensibilidad



Error, Exactitud y Precisión

- *Si no se lo mide, no se lo puede administrar*
- Tiene implicancias:
 - Sobre el dato final
 - Sobre el proceso productivo
 - Sobre las habilidades de las personas
- Cada dato es específico, pero hay reglas generales

Definiendo estándares

Fuente: Foote and Huebner

- Especificar criterios desde el principio
- Para datos espaciales y no espaciales
- Tópicos a considerar:
 - Niveles de error admisible
 - Diccionarios, tesauros
 - Criterios de clasificación
- Criterios para Datos
- Criterios para Procedimientos
- Tres pasos:

1: qué exigir

- Especificar criterios
 - Mucha exactitud *cuesta*
 - Poca exactitud *cuesta*
 - Requisitos son propios del proyecto
- Ir hacia atrás: desde el output a los inputs
- Requerirán Análisis de Sensibilidad

2: entrenar a la gente

- No alcanza con exigir el éxito
- Deben conocerse los objetivos
- Debe capacitarse para lograrlo
- Deben saber *para qué* se hace así

El proyecto no puede salir *a pesar* de la gente

3: verificar procesos y resultados

- Aplique un control regular
- El Control puede ser
 - por lotes
 - continuo
- Aplicarlos
 - A proveedores externos
 - Internamente



Plan

- ✓ Error, Exactitud y Precisión
- ✓ Definiendo estándares
- Informando sobre la Calidad
- Midiendo y Verificando
- Análisis de Sensibilidad



Informando sobre la Calidad

- Siempre debe estar escrito
 - En documentos independientes
 - Dentro de los datos
- Incluir Cómo, Cuándo, Dónde
- Asignar a un responsable
- Documentar preserva la inversión
 - Los datos sobreviven más de lo esperado
 - Típico es 50-100 años
- Y después vendrán los arqueólogos...

Plan

- ✓ Error, Exactitud y Precisión
- ✓ Definiendo estándares
- ✓ Informando sobre la Calidad
- Midiendo y Verificando
- Análisis de Sensibilidad



Midiendo y Verificando

- La realidad manda...
- Comparar en campo si es posible
- Comparar contra otra fuente más precisa si existe
- Existen ensayos específicos por tipo de dato
- Se aplican tests paramétricos y no paramétricos

Cualquiera sea el test...

- Deberá ser científicamente creíble
 - Basado en estadística, medidas o ambas
 - Repetible por otros
- Deberá costar algo razonable
 - Costo en dinero
 - Costo en tiempo
- Ampliamente aceptable
 - Basado en estándares profesionales, nacionales y/o internacionales

Veremos uno...

NSSDA: National Standard for Spatial Data Accuracy (1998)

NSSDA

POSITIONAL ACCURACY HANDBOOK 1999

- Identifica un estadístico bien definido para describir resultados de ensayos de exactitud posicional
- Describe un método para el ensayo de coordenadas o posiciones
- Provee un lenguaje estandarizado para informar la exactitud
- No prevee ensayos para variables temáticas o categóricas; sólo cuantitativas

Siete pasos

1. Determinar si se requiere exactitud horizontal, vertical o ambas
2. Seleccionar un conjunto de test points
3. Identificar y extraer los homólogos de un conjunto independiente de mejor exactitud
4. Tomar medidas en puntos idénticos a los anteriores

Siete pasos (cont)

5. Calcular un estadístico de la exactitud de la posición según corresponda
6. Preparar la declaración de exactitud según el formulario
7. Incluya esa declaración en los metadatos



Siete pasos

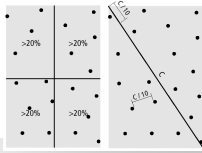
1. Determinar si se requiere exactitud horizontal, vertical o ambas
2. Seleccionar un conjunto de test points
3. Identificar y extraer los homólogos de un conjunto independiente de mejor exactitud
4. Tomar medidas en puntos idénticos a los anteriores

Siete pasos

1. Determinar si se requiere exactitud horizontal, vertical o ambas
2. Seleccionar un conjunto de test points
3. Identificar y extraer los homólogos de un conjunto independiente de mejor exactitud
4. Tomar medidas en puntos idénticos a los anteriores

Algunos detalles...

- ¿Dónde localizar los puntos?
- Bien definidos, fáciles de medir
- Intersecciones perpendiculares
- Monumentos
- $N > 20$, fallas ≤ 1
- $N \leq 20$?
 - Estimación deductiva
 - Evidencia interna
 - Comparación con la fuente



Siete pasos

1. Determinar si se requiere exactitud horizontal, vertical o ambas
2. Seleccionar un conjunto de test points
3. Identificar y extraer los homólogos de un conjunto independiente de mejor exactitud
4. Tomar medidas en puntos idénticos a los anteriores

Datos ¿independientes?

- Buscar puntos comunes
- Exactitud
 - Idealmente >> triple
 - En otro caso... lo que haya (¡si es mejor!)
- Documentarla en metadatos
- ¡Usar sólo cifras significativas!
- Exactitud ~uniforme en área
- Área ~equivalente

Siete pasos (cont)

- Calcular un estadístico de la exactitud de la posición según corresponda
- Preparar la declaración de exactitud según el formulario
- Incluya esa declaración en los metadatos



Cálculos en horizontal...

$$RMSE_x = \sqrt{\frac{\sum_{j=1}^n (x_{data,j} - x_{check,j})^2}{n}}$$

$$RMSE_y = \sqrt{\frac{\sum_{j=1}^n (y_{data,j} - y_{check,j})^2}{n}}$$

$$\left(\frac{RMSE_x}{RMSE_y} \in \left[\frac{6}{10}, \frac{10}{6} \right] \right)$$

Si: { Error_x independiente de Error_y
Errores con distribución normal
No hay outliers

$$\rightarrow \text{Exactitud}_{\text{NSSDA,R}} = 2.4477 * 0.5 * (RMSE_x + RMSE_y)$$

Cálculos en la vertical...

$$RMSE_z = \sqrt{\frac{\sum_{j=1}^n (z_{data,j} - z_{check,j})^2}{n}}$$

Si: $\begin{cases} \text{Errores con distribución Normal} \\ \text{No hay outliers} \end{cases}$

$$\rightarrow \text{Exactitud}_{NSSDA,Z} = 1.9600 * RMSE_Z$$

Variables temáticas o categóricas

- No contempladas en el NSSDA
- Es importante hacer foco en:
- Naturaleza de los errores
 - ¿Se confundieron especies de pinos?
 - ¿Se confundió trigo con vid?
 - Frecuencia de los errores
 - Magnitud de los errores
 - Origen de los errores

Algunos serán más graves que otros

¿Cómo medir la Exactitud?

- Prepare una matriz de confusión
 - Elija N puntos al azar
 - Vea lo que dicen los datos
 - Vea lo que hay en el terreno
 - Cuente los casos
 - Idealmente, todo en la diagonal
- | | | | |
|-------|---|----|---|
| | | | T |
| | | A | |
| Datos | A | 10 | |
| | B | 0 | |

		Terreno		
		A	B	C
Datos	A	10	2	3
	B	0	20	0
	C	4	1	10

Exactitud global=traza(A)/N

Alguna terminología

- **Error por Omisión**
Def.: Error por columna/total en columna
(1.0 - Exactitud *del productor*)
- **Error por Comisión**
Def: Error por fila/total en fila
(1.0-Exactitud *del usuario*)

¡No incluir los elementos de la diagonal!

Resultados del ejemplo

- Exactitud *Total*:
 $(10+20+10)/(10+2+3+0+20+0+4+1+10) = 40/50 = 80\%$
- Exactitud de *comisión* para la clase A:
 $10/(10+2+3) = 10/15 = 67\%$
- Exactitud de *omisión* para la clase A:
 $10/(10+0+4) = 10/14 = 71\%$

Datos	Terreno		
	A	B	C
A	10	2	3
B	0	20	0
C	4	1	10

La tabla completa

	A	B	C	TOTAL	Exactitud del Usuario
A	10	2	3	15	0.67
B	0	20	0	20	1.00
C	4	1	10	15	0.67
TOTAL	14	23	13	40	
Exactitud del Productor	0.71	0.87	0.77		
Exactitud total	0.8				

El índice *kappa* de Cohen

- Una medida de la exactitud observada en comparación con el azar
- Es un único número (un escalor)
- El ideal sería 1.0; lo peor sería 0.0
- Un valor intermedio 0.83 implica que se están evitando 83% de los errores que cometería una clasificación al azar
- No está recogido en los estándares
- Sin embargo se usa...
- ¿Cómo calcularlo?

Ejemplo:

	A	B	C	TOTAL	Exactitud del Usuario
A	10	2	3	15	0.67
B	0	20	0	20	1.00
C	4	1	10	15	0.67
TOTAL	14	23	13	40	
Exactitud del Productor	0.71	0.87	0.77		
Exactitud total	0.8				

$$q_A = 15 \cdot 14 / 50 = 4.20$$

$$q_B = 20 \cdot 23 / 50 = 9.20$$

$$q_C = 15 \cdot 13 / 50 = 3.90$$

$$q = \text{Total} = 17.3$$

1. Calcular "q" para cada caso
 Def.: q es el número de casos en la diagonal por puro azar
 $q_i = \text{sum}_i(\text{fila}) \cdot \text{sum}_i(\text{col}) / N$

Ejemplo (cont.):

	A	B	C	TOTAL	Exactitud del Usuario
A	10	2	3	15	0.67
B	0	20	0	20	1.00
C	4	1	10	15	0.67
TOTAL	14	23	13	40	
Exactitud del Productor	0.71	0.87	0.77		
Exactitud total	0.8				

$$\text{Kappa} = (40 - 17.3) / (50 - 40) = 0.227$$

2. Calcular *kappa*

$$\text{Def.: } \text{kappa} = (\text{traza}(A) - q) / (N - q)$$

¡Bastante malo!

Un pequeño problema...

- Se asume que no hay confusión posible al construir la matriz de confusión
- En la práctica...
 - Absolutamente incorrecto
 - Comprensible, pero incorrecto
 - Razonable, pero podría ser mejor
 - Buena respuesta
 - Absolutamente correcto

Digresión: Midiendo Calidad...

- Estamos revisando los siete pasos
- Hemos visto algunos tests de Exactitud
- Algunos estandarizados, otros no
 - Para posición
 - Para atributos
- Esencialmente controlan Exactitud
- Pero en Calidad no todo es Exactitud...

Otras componentes...

- Completitud
- Coherencia lógica
- Exactitud temporal
- Linaje
- Metadatos

Componentes más modernas
 Sin definir totalmente
 Falta unificar nomenclatura
 Faltan métricas

Ejemplo: Coherencia lógica

- Representación sin sentido
 - Caminos en el agua
 - Isolinneas que se cortan
 - Puentes sin rutas o sin río
 - Cuencas hidrográficas vs. ríos



¿Cómo lo mediría?

Volviendo a los Siete pasos...

5. Calcular un estadístico de la exactitud de la posición según corresponda
6. Preparar la declaración de exactitud según el formulario
7. Incluya esa declaración en los metadatos



Declaración de Exactitud

POSITIONAL ACCURACY HANDBOOK 1999

- En el estándar se consideran dos formatos:
 1. "___ mts de exactitud (horizontal/vertical) al 95% de confianza"
 2. "Compilado para tener ___ mts de exactitud (horizontal/vertical) al 95% de confianza"
- Usar (1) si se dispone de datos específicos independientes
- Usar (2) si se utilizó un procedimiento estándar consistente con el error obtenido

Incluyendo la Incertidumbre

- Falacia de *Falsa Precisión*
 - Una represa de 1:234.567,89 m³
 - La probabilidad de falla es 0.3579
- Regla (indefendible...): un décimo de la resolución
 - Población en unidades → densidad en décimos
 - Simple y popular, ¡pero errónea!

Incluyendo Incertidumbre⁽²⁾

- Falacia de *Falsa Certidumbre*
 - Respuestas "precisas" sin sustento
 - Exacerbado al combinar capas
 - Ni hablar de mezclar escalas
 - Problema extremadamente corriente
- Lo correcto sería alguno de:
 - resultado+intervalo
 - [peor, típico, mejor]
 - [min,max] con probabilidad
- ¿Cómo determinarlo?

¡Permanezca en este canal!

Siete pasos (cont)

5. Calcular un estadístico de la exactitud de la posición según corresponda
6. Preparar la declaración de exactitud según el formulario
7. Incluya esa declaración en los metadatos



Comunicando...

POSITIONAL ACCURACY HANDBOOK 1999

- Si hay varias capas, podría requerirse una declaración para cada una de ella
- Si hay una mezcla inseparable y no ensayada, asignarle el peor nivel de error
- Si está ensayada, asignarle el nivel obtenido
- Nunca use más cifras que las de los datos

Otras formas de estimar...

- Si tiene pocos puntos
- Si la distribución es claramente no-gaussiana
- ¡No está previsto en el estándar!
 - ➔ Use Remuestreo (*Bootstrap*)

Atención: Lo que de allí resulte complementa, pero no sustituye a lo requerido por el estándar

Breve introducción al Bootstrap

¿Porqué Remuestrear?

Fuente: Lucila Ohno-Machado

- En ocasiones no es posible tener muchas muestras de una población
- En ocasiones no es correcto (o posible) asumir una distribución para una población
- El objetivo es: Assess sampling variation

Bootstrap

- Efron (biostadístico de Stanford) a finales de los 80's
 - "Pulling oneself up by one's bootstraps"
- Es un enfoque no paramétrico para inferencia estadística
- Usa *cálculos (fuerza bruta)* en lugar de resultados asintóticos e hipótesis tradicionales sobre distribuciones
- Puede usarse para obtener errores estándar, intervalos de confianza y prueba de hipótesis

Ejemplo

- Adaptado de Fox (1997) "Applied Regression Analysis"
- Objetivo: Estimar diferencia promedio de respuestas entre Hombre y Mujer
- Se dispone únicamente de cuatro parejas de observaciones:

Observ.	Hombre	Mujer	Differ.
1	24	18	6
2	14	17	-3
3	40	35	5
4	44	41	3

Diferencia promedio

- Promedio \bar{Y} de la muestra es
 $(6+3+5+3)/4 = 2.75$
- Si Y fuese normal, el Intervalo de Confianza CI al 95% sería

$$\mu = \bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$
- Problema: no conocemos σ

Estimadores corrientes

- El estimador de σ es

$$S = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{(n-1)}}$$
- El estimador de la desviación estándar es

$$SE(\bar{Y}) = \frac{S}{\sqrt{n}}$$
- Asumiendo que la población es normal, podemos usar la distribución-t como

$$\mu = \bar{Y} \pm t_{n-1, 0.025} \frac{S}{\sqrt{n}}$$

Intervalo de Confianza CI

$$\mu = \bar{Y} \pm t_{n-1, 0.025} \frac{S}{\sqrt{n}}$$

$$\mu = 2.75 \pm 4.30 (2.015) = 2.75 \pm 8.66$$

$$-5.91 < \mu < 11.41$$

¡¡¡¡¡¡¡¡¡¡

Media y varianza de la muestra

Idea: usar la distribución Y^* de la muestra para estimar la distribución Y de la población

y^*	$p^*(y^*)$	
6	.25	
-3	.25	$E^*(Y^*) = \sum y^* p(y^*) = 2.75$
5	.25	$V^*(Y^*) = \sum [y^* - E^*(Y^*)]^2 p(y^*)$
3	.25	$= 12.187$
		$DE^*(Y^*) = \sqrt{V^*(Y^*)} = 1.745$

Muestreo con reemplazo

Muestra	Y_1^*	Y_2^*	Y_3^*	Y_4^*	\bar{Y}^*
1	6	6	6	6	6.00
2	6	6	-3	3	3.75
3	6	6	5	5	5.75
...					
100	-3	5	6	3	2.75
101	3	3	3	3	3.00
...					
255	-3	3	3	5	3.5
256	3	3	3	3	3.00

Caso extremo

Sólo hay 256 posibilidades

Otro caso extremo

Calculando nuevos CI

El promedio de estos 256 promedios es nuevamente 2.75, pero su desviación es ahora

$$DE^*(\bar{Y}^*) = \sqrt{\frac{\sum (\bar{Y}_k^* - \bar{Y}^*)^2}{n^*}} = 1.745$$

(sin "n" porque la desviación ahora no se estima)

$$DE(\bar{Y}) = \sqrt{\frac{n}{n-1}} DE^*(\bar{Y}^*)$$

$$2.015 = \sqrt{\frac{4}{3}} \times 1.745$$

¡¡¡¡¡¡¡¡¡¡

¿Y entonces?

■ ¡Esto ya era sabido!

■ Pero mediante remuestreo o bootstrap

- Los Intervalos de Confianza pueden ser más precisos
- Pueden ser calculados para problemas no lineales que no tienen fórmulas de error conocidas

**La población es a la muestra
lo mismo que
la muestra es a las remuestras**

En la práctica (a diferencia del ejemplo anterior), no *todas* las N^N remuestras son seleccionadas

Procedimiento

1. Especifique un criterio que produzca la muestra que utilizará

Criterio(población) → muestra

2. Use esta muestra como si fuese la población (pero *con reemplazo*)

Criterio(muestra) → re-muestra1
re-muestra2
etc...

Léase como "el criterio aplicado a la población"

Procedimiento⁽²⁾

3. Para cada remuestra, calcule el estimador estadístico de su interés
4. Use la distribución de los estimadores de las remuestras para estimar las propiedades de la muestra

Atención:

- Remuestreo no es válido *absolutamente* para todo
- No debe usarse si hay colas "largas"
- Ej.: estimación del rango

<http://www.itl.nist.gov/div898/handbook/eda/section3/boottplot.htm>

Otro ejemplo:

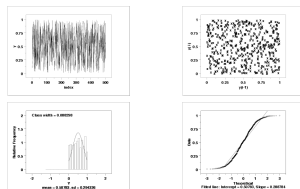
- Ver si unos datos ajustan al modelo

$$Y_i = C + e_i$$

- Confirmar que el proceso está *bajo control*
 - Tomar muestras al azar y confirmar que es de distribución, parámetros y escala fijos
- En particular, confirmar CI al 95%

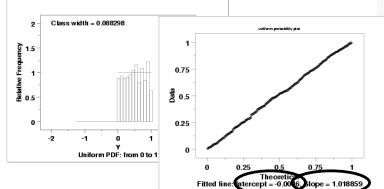
$$CI = \left[\bar{Y} - 2.0 \frac{s}{\sqrt{n}}, \bar{Y} + 2.0 \frac{s}{\sqrt{n}} \right]$$

¿Qué hacer primero? 4 gráficos



No luce como normal...

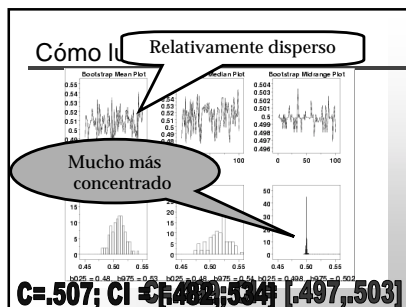
Supongamos Uniforme...



Ahora sí...

Un estimador para C

- Y_i no es normal
- Aplicaremos remuestreo
- Estimadores posibles de C
 - Media
 - Mediana
 - Promedio del rango
 - Quizá otros...
- Buscaremos el de mínima varianza



Importancia...

- Lo típico es requerir CI para la media
 - Caso normal: hay solución analítica
- En otros casos no la hay
 - Es donde el bootstrap/resampling muestra su importancia

Otros métodos de remuestreo

- Jackknife (*sacar de a uno*) es un caso especial de bootstrap
 - Remuestreo sin un caso y sin reemplazo (las muestras pasan a tener tamaño $n-1$)
- Validación cruzada
 - Divide los datos en entrenamiento y test
- Generalmente se los usa para estimar intervalos de confianza en predicciones para el modelo "full" (i.e., modelo que usa todos los casos)

Recapitulando:

- Vimos estándares
- Vimos métricas de Exactitud
 - Posicionales
 - Atributos Cuantitativos, Cualitativos, etc
- Vimos métodos para medir Exactitud
 - Cuando hay acceso a *suficientes* valores *mejores*
 - Cuando son gaussianos
- Para los otros casos
 - Explicamos e ilustramos Bootstrap
 - Mencionamos Jackknife

Plan

- ✓ Error, Exactitud y Precisión
- ✓ Definiendo estándares
- ✓ Informando sobre la Calidad
- ✓ Midiendo y Verificando
- Análisis de Sensibilidad

Análisis de Sensibilidad

- Siempre hay que Validar los resultados
 - Conjunto de entrenamiento
 - Conjunto de validación
- ¿Cómo? Depende del Modelo+Datos
- Datos continuos → derivada parcial
- Datos categóricos → simulación
- Se suele hacer al final... ¡pero puede invalidar todo el proyecto!

Cuanto antes mejor...

- Si el dato no existe con la exactitud requerida
- Si no era necesaria tanta exactitud ¡Todo cuesta!
- Datos pueden ser irrelevantes...
 - Tipo de suelo para estudio eólico
- Relevancia diferente según el output
 - Producción de energía en kWh-año
 - Costo del proyecto en US\$

Ejemplo analítico 1

- Supongamos $f(x) = ax + b$
- ¿Cómo se relaciona la Var(f) con Var(x)?

$$\begin{aligned}
 V(f) &= \langle f^2 \rangle - \langle f \rangle^2 \\
 &= \langle (ax+b)^2 \rangle - \langle ax+b \rangle^2 \\
 &= a^2 \langle x^2 \rangle + 2ab \langle x \rangle + b^2 - a^2 \langle x \rangle^2 - 2ab \langle x \rangle - b^2 \\
 &= a^2 \langle x^2 \rangle - \langle x \rangle^2 \\
 &= a^2 V(x) \quad \text{ i.e. } \sigma_f = |a| \sigma_x
 \end{aligned}$$
- Más en general:

$$V(f) = \left(\frac{df}{dx} \right)^2 V(x) \quad ; \quad \sigma_f = \left| \frac{df}{dx} \right| \sigma_x$$

Pero vale sólo si la *aproximación lineal* es buena en el rango del error

Ejemplo analítico 2

- Consideremos ahora $f = ax + by + c$

$$\begin{aligned}
 V(f) &= a^2 \langle x^2 \rangle - \langle x \rangle^2 + b^2 \langle y^2 \rangle - \langle y \rangle^2 + 2ab \langle xy \rangle - \langle x \rangle \langle y \rangle \\
 &= a^2 V(x) + b^2 V(y) + 2ab \text{cov}(x, y)
 \end{aligned}$$
- En general

$$\begin{aligned}
 V(f) &= \left(\frac{df}{dx} \right)^2 V(x) + \left(\frac{df}{dy} \right)^2 V(y) + 2 \left(\frac{df}{dx} \right) \left(\frac{df}{dy} \right) \text{cov}(x, y) \\
 \sigma_f^2 &= \left(\frac{df}{dx} \right)^2 \sigma_x^2 + \left(\frac{df}{dy} \right)^2 \sigma_y^2 + 2 \left(\frac{df}{dx} \right) \left(\frac{df}{dy} \right) \rho \sigma_x \sigma_y
 \end{aligned}$$

El coeficiente de correlación ρ es 0 si x,y no están correlacionados

Nuevamente sólo es válida si la aproximación lineal es buena en el rango del error

Otras fórmulas analíticas

- Ahora considérese $f = x \cdot y$ (solo creámoslo...)
 $V(f) = y^2 V(x) + x^2 V(y)$

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$$

➤ Resultados similares para $f = x/y$
- Otras fórmulas útiles
 $\frac{\sigma_{1/x}}{1/x} = \frac{\sigma_x}{x}$; $\sigma_{\ln(x)} = \frac{\sigma_x}{x}$
El error relativo en x o 1/x es el mismo Error en el logaritmo es el error relativo

En general...

- Hay que estimar las derivadas parciales
- Métodos numéricos *aproximados*:
 - Cociente incremental

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x_1, x_2, \dots, x_i + \Delta, \dots, x_n) - f(x_1, x_2, \dots, x_i, \dots, x_n)}{\Delta}$$
 - Uso de números complejos

$$\frac{\partial f}{\partial x_i} = \lim_{\Delta \rightarrow 0} \frac{\text{Im}(f(x_1, x_2, \dots, x_i + \Delta \cdot j, \dots, x_n))}{\Delta}, j^2 = -1$$
- Problema: hay que estimar incremento Δ apropiado
- Alternativa: uso de la derivada *exacta*

Derivada exacta...

- Derivación manual
 - Sólo en casos simples; require codificar
- Derivación automática
 - Variante 1: Generador de nuevo código
 - Variante 2: Sobrecarga de operadores
- Paquetes
 - ADOL-F/ADOL-C
 - TAPENADE
 - Matlab+ADMAT
 - Otros...
- Ejemplos



Generadores de código

- Generan automáticamente un segundo código fuente a partir del disponible
- ADOL-C/ADOL-F/TAPENADE, etc.
- Derivada exacta; no hay cociente incremental
- Ejemplo:

$$\begin{aligned} v_2 &= 2 * v_1 + 5 \\ \dot{v}_2 &= 2 * \dot{v}_1 \\ v_4 &= v_2 + p_1 * v_3 / v_2 \\ \dot{v}_4 &= \dot{v}_2 * (1 - p_1 * v_3 / v_2^2) + \dot{v}_3 * p_1 / v_2 \end{aligned}$$

Sobrecarga de operadores

- Sólo en ambientes orientados a objetos
- Ej: Matlab + ADMAT toolbox


```
function y = getfun(x)
    z = x*x;
    w = x + z;
    y = w*w
```

$x = 2.0$
 $\dot{x} = 1.0$

$z = 4.0$
 $\dot{z} = 2 * x * \dot{x} = 4.0$

$w = 6.0$
 $\dot{w} = \dot{x} + \dot{z} = 5.0$

$y = 36.0$
 $\dot{y} = 2 * w * \dot{w} = 60.0$
- Ventajas:
 - Derivada exacta (tampoco hay cociente incremental)
 - Sólo hay un código fuente

¿En qué andábamos?

- Estábamos haciendo un Análisis de Sensibilidad
 - ¿Alcanza/no alcanza la Exactitud de los datos disponibles?
- Vimos Métodos Analíticos
 - Sólo para datos continuos
- Veremos métodos Estadísticos de Simulación
 - Para datos continuos/categoricos



Métodos de Monte Carlo

Fuente: Vesna Luzziar-Stiffler

- Cualquier procedimiento que usa números aleatorios
- ¿Números aleatorios?
 - ¿Qué es un número aleatorio?
 - ¿Un número?
 - Secuencia de números aleatorios



La parte aleatoria...

- Distribución Uniforme
 - Todo número en $[0,1]$ tiene la misma probabilidad
 - Fundamental (se usa para otras distribuciones)
 - Para muestreo al azar, etc.
- ¿Cómo generar una secuencia de números aleatorios?
 - Usar un sistema caótico (moneda, dado, etc.)
 - Usar un proceso inherentemente randómico
 - Usar tablas de números aleatorios
- Problema:
 - Resultados no reproducibles
 - Impráctico, no confiable (¿sesgos?)
 - Tablas - no suficientemente extensas

Números pseudo-aleatorios

- Números pseudo-aleatorios generados con algoritmos
 - No correlacionados, ciclos largos, etc.
- Ej: Middle square algorithm (J. Von Neumann, 1946):
 - Para generar una secuencia de enteros de 10 dígitos:
 - Elige uno cualquiera,
 - Elevalo al cuadrado y luego
 - Extraiga los 10 dígitos intermedios como el siguiente número de la secuencia
 - Por ejemplo:

$$3690295441^2 = 13618280441885334481$$

Seudo-Aleatorios...

- ¡No son al azar! Son series totalmente determinísticas y predecibles
- Lucen como *al azar*, pero ...

$$\begin{array}{r} 6100^2 \\ 3721000 \\ 4410000 \\ 14610000 \\ 65610000 \end{array}$$
 ¡LOOP!

Linear congruential method (Lehmer, 1948)

- $I_{n+1} = (a \cdot I_n + c) \bmod m$
 - I_0 = Valor inicial (semilla)
 - $a, c \geq 0$,
 - $m > I_0, a, c$
- Una mala elección de constantes generará secuencias malas:
 - Ej.: $a=c=I_0=7, m=10 \rightarrow$

$$\{7, 6, 9, 0, 7, 6, 9, 0, \dots\}$$

A mod B=resto de dividir A entre B

Familia popular...

- RANDU: $I_{n+1} = (65539 \cdot I_n) \bmod 2^{31}$
 - Se vio que no era bueno
- RANMAR: $I_n = (a \cdot I_{n-1} + b \cdot I_{n-2}) \bmod m$
 - genera series con periodo $\approx 10^{43}$
- En general se deberían usar generadores conocidos, con propiedades bien documentadas
- Sin embargo...

Para PDF's no uniformes

- Puede haber rutinas específicas (ej.: $N(0,1)$)
- ¿Otros casos?
 - Método de la probabilidad inversa
 - Calcular $F(x) = \int_{-\infty}^x f(s) ds$
 - Generar $t \sim U(0,1)$
 - Evaluar $x = F^{-1}(t)$
 - x resulta tener pdf igual a $f(x)$
 - Hay otros métodos para casos particulares
- Problemas: pdf's discretas, eficiencia, etc.

Simulación de Monte Carlo

- Aparece con las primeras computadoras (1945-55)
- Áreas de aplicación incluyen biología, química, informática, análisis de datos econométricos y financieros, ingeniería, ciencia de los materiales, física, ciencias sociales, estadística, etc.
- La SMC se resume en:
 - Asumiendo un mecanismo para generar datos de su proceso
 - Produzca nuevas instancias de datos simulados
 - Examine estadísticamente los resultados de esas instancias
- Si se tiene "la" pdf \rightarrow SMC estándar
- Si no, se usa Remuestreo (que es un pariente cercano)

Típicamente medias y varianzas

Algunos problemas...

- SMC plantea gran demanda de CPU
 - Varianza $\sim 1/\sqrt{n}$
 - Bajar el error un 10% requiere multiplicar por 100 el número de simulaciones
- Se intenta mejorar esta característica
 - Técnicas de Reducción de Varianza (TRV)
- Veamos de qué se trata...

TRV: Enfoque Antitético

- Se basa en que

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$
- Si X_1 y X_2 son independientes,

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$
- Si se pueden generar X_1 y X_2 de forma que tengan correlación $\text{Cov}(X_1, X_2)$ negativa puede lograrse que $\text{Var}(X_1 + X_2)$ sea *menor* que si son independientes

Ejemplo: Integral (o promedio)

$$\hat{I} = \frac{1}{2n} \sum_{i=1}^{2n} X_i \quad \text{donde } X_1, X_2, \dots, X_{2n} \stackrel{\text{iid}}{\sim} f(x)$$

Puede ser reformulado como

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{donde } Y_i = \frac{X_i + X'_i}{2}$$

donde X_i y X'_i están correlacionados negativamente y las Y_i siguen siendo independientes

¿Cómo lograr correlación negativa?

- El procedimiento en general es

$$X = F^{-1}(U) \quad \text{donde } U \sim U(0,1) \text{ y } dF/ds = f(s)$$
- Pero $U \sim U(0,1)$ implica que $(1-U) \sim U(0,1)$.
- Por lo que, $X' = F^{-1}(1-U)$ también tiene la misma distribución $f(s)$ que X
- Si $F(s)$ es monótona (y por lo tanto también $F^{-1}(s)$) entonces X' y X tienen correlación negativa

Otros trucos...

- Si se quiere calcular $I = \int_a^b g(x) dx$
- En ocasiones es fácil resolver analíticamente la integral $I' = \int_a^b h(x) dx$
- y posteriormente hacer la simulación de Monte Carlo en el residuo, logrando

$$I = I' + \int_a^b [g(x) - h(x)] dx$$

MonteCarlo

Aspectos importantes en SMC

- Diseño apropiado
- Medida de la exactitud del resultado
- Justificación del número de casos/muestras
- Selección del generador aleatorio
- Memoria y tiempo de cálculo
- Técnica de reducción de varianza
- Software usado
- Análisis del resultado (visualización, ajuste por curvas, etc.)



En resumen:

- Como productor: hay estándares para expresar Exactitud
 - Para posición y variables continuas
 - No para variables categóricas
- Estimación basada en suficientes valores más exactos
 - Alternativamente, usar Bootstrap
- Como usuario: preocuparse *mucho* sólo si el problema es *sensible*

¿Cómo medir Sensibilidad?

- Hay técnicas informáticas...
 - Simulación Monte Carlo
 - Uso general, no paramétrica, etc.
 - Mucha CPU y tiempo
 - Propagación de errores
 - Sólo funciones diferenciables
 - No requiere simulación
 - Complejas para algunas operaciones GIS típicas
- Escasa disponibilidad en GIS

Módulo 3: Midiendo la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org

Módulo 4: Mejorando la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org

Plan

- Introducción
- Revisión de herramientas estadísticas
- Detectando problemas
- Imputando valores ausentes
- Ejemplos



Condicionantes...

Éxito depende de:

- Disponibilidad de Datos
- Disponibilidad de Modelos
- Sensibilidad de los Modelos
- Capacitación de técnicos
- Calidad de Datos
- Otros



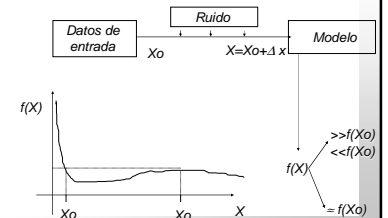
Disponibilidad de Datos

- ¡Siempre limitada!
 - Atributos
 - Resolución espacial
 - Vigencia
 - Niveles de Error
 - Otros... (\$, inexistencia)
- Normalmente ¡condicionan al modelo!

Disponibilidad de Modelos

- Modelo no es lo mismo que Realidad
- Siempre imperfectos
 - Quizá "importados" de USA, etc. ☹
 - Suelen faltar datos
 - Datos sustitutivos (más otros modelos...)
 - Poco plazo, poco presupuesto...
- Usualmente no validados
- Códigos complejos (CPU, disco, etc.)

Sensibilidad del Modelo



Sensibilidad del Modelo⁽²⁾

- Es específica al conjunto {Modelo,Datos}
- Un problema en si mismo
 - ¿Qué tipo de errores? ¿Cuántos? ¿Dónde?
 - Enfoque Determinístico
 - Enfoque Estocástico
- Ejs.: Viewshed area (Fisher)
- Ejs.: Goodchild para líneas

Capacitación de los técnicos

- Idealmente deberían:
- Conocer del problema "físico"
 - Conocer de los datos (propios y ajenos!)
 - Conocer los modelos
 - Capaces de criticar resultados



¡Es mucho conocer!

Condicionantes...

Éxito depende de:

- ✓ Disponibilidad de Datos
- ✓ Disponibilidad de Modelos
- ✓ Sensibilidad de los Modelos
- ✓ Capacitación de técnicos
- Calidad de Datos
- Otros



Calidad de Datos

- Completitud
 - Exactitud
 - Vigencia
 - Linaje
- Si no son "apropiados":
- Buscar fuentes alternativas
 - Arremangarse...
 - Mejorar Exactitud
 - Cambiar de Modelo



Dos actores...

- Usuario:
 - Tomador de Datos
 - Sufridor de Consecuencias
 - +productos, con -fondos
- Productor:
 - Receptor de Críticas
 - Usualmente monopolico
 - +productos, con -fondos

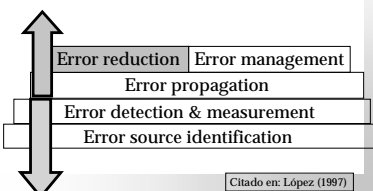
¡Me pesimista!

Dos actores... (versión optimista)

- Usuario:
 - Especifica requerimientos
 - Preocupado por la Exactitud
 - No tiene acceso fluido a "la verdad"
 - Llevará la Culpa...
- Productor:
 - Observa estándares
 - Preocupado por la Exactitud
 - La "verdad" existe, pero es más cara
 - Llevará la Culpa...

¡El problema común!

Una jerarquía de necesidades...



Fuera de discusión (S, plazo de entrega, etc.)

¿Problema para algún PhD?

Conocimiento insuficiente de las relaciones cuantitativas

Carencia de datos apropiados e independientes para validar

Conocimiento insuficiente de la sensibilidad del modelo

¿Dónde están los outliers que importan?

¿Cómo imputar los valores ausentes?

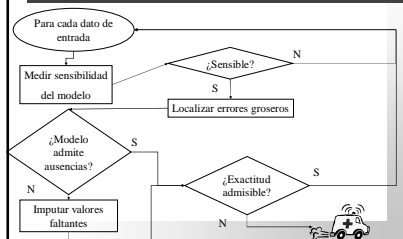
¡lo posible...

El proceso requeriría...

- Evaluar la sensibilidad del modelo
- Localizar errores groseros (outliers)
- Asignar valores apropiados para los outliers y/o los faltantes

¡Casi nada de ello está previsto en un GIS corriente!

Diagrama de decisión



Análisis de Sensibilidad

- No pueden analizarse *todos* los modelos
- Ejemplos:
 - Modelo hidrológico de una cuenca
 - Inputs: lluvia diaria, caudal en ríos, uso del terreno, geología, etc.
 - Outputs: caudal en ríos, niveles en las presas
 - Modelo de contaminación de aire
 - Inputs: inventario de emisores, viento de superficie, MDE, etc.
 - Outputs: mapas de niveles de inmisión

¿Sensibilidad...? ¿Cómo?

- Propagación analítica
 - Taylor
 - Aritmética de Intervalos
- Monte Carlo

Temas:

- ¿Generación de "errores"?
 - Tamaño, localización, correlación...
- ¿Generación de outliers?
 - ¿PDF?, ¿modelo del error?

Fuente: Burrough & McDonnell (1998)

Expansión de Taylor

- Nos restringimos a modelos que son función del punto x *(para facilitar, think raster)*
 - Excluimos buffers, ventanas, topologías, etc.
- Si el modelo puede ponerse como
$$U = g(A_1, A_2, \dots, A_n)$$
siendo A_i atributo cuantitativo sujeto a error
- Se conocen $\langle A_i \rangle$ y $\text{var}(A_i)$; también $\text{var}(A_i, A_j)$
- Si $g(\cdot)$ es lineal, entonces es fácil

Propagación en el Caso lineal

Si
$$U = g(A_i) = \sum_{i=1}^n b_i A_i$$
y los A_i no están correlacionados, entonces
$$\langle U \rangle = \sum_{i=1}^n b_i \langle A_i \rangle$$
y
$$\text{var}(U) = \sum_{i=1}^n b_i^2 \text{var}(A_i)$$
Si hay correlación, entonces
$$\text{var}(U) = \sum_{i=1}^n \sum_{j=1}^n b_i b_j \sqrt{\text{var}(A_i) \text{var}(A_j)} \rho(i, j)$$

Caso más general

- Linealiza la función $g(\cdot)$
- Taylor al primer orden
$$\delta U = \sum_{i=1}^n \frac{\partial g}{\partial A_i} \delta A_i + \dots$$
- ¡Equivale a una función g lineal! ➔ caso conocido
- Algunos autores llegan hasta segundo orden
- O dicen que llegan... ☺

Pros y Contras

- Ventajas:
 - Es una fórmula analítica
 - Eficaz en términos de CPU
 - Maneja correlación espacial
- Problemas: se trata de una *aproximación*
 - ¿Será buena? ¿mala?
 - ¿De dónde saco las derivadas parciales?
 - Es fácil si hay normalidad $N(0, \sigma)$
 - En algunos casos el error no tiene media cero
 - ¿Cómo estimar la correlación espacial del error?

Cálculo de derivadas parciales

- A mano, cualquiera *podría*
 - Sólo modelos chicos, relativamente simples
- Soluciones de hoy
 - Álgebra simbólica (Maple, Derive, etc.)
 - Procesadores de Código fuente
 - ADOL-C/ADOL-F
 - Tapenade
 - Sobrecarga de operadores
 - Matlab+ADMAT
 - C++, F90, etc.

Eso no es todo...

- En general, los errores son *función de punto* y no constantes espaciales
 - Ej.: interpolación
- Eso afecta a la estimación de δA_i
- El procedimiento estándar es Kriging
 - ¡Pero Kriging no genera outliers!
- ¿Cómo generar errores groseros?
 - Yet to be solved...

Fuente: B. Schneider

Aritmética de Intervalos

- También analítico
- Equivalente a un "peor caso"
- Notación: Si $a_i \leq A_i \leq \bar{A}_i \Rightarrow [A_i] = [a_i, \bar{A}_i]$
- Ej.:
 - Suma: $S = A_i + B_i$; $[S] = [a_i + b_i, \bar{A}_i + \bar{B}_i]$
 - Producto: $P = A_i \cdot B_i$;
 $[P] = [\min(a_i b_i, a_i \bar{B}_i, \bar{A}_i b_i, \bar{A}_i \bar{B}_i), \max(\text{idem})]$
- Automatizable
 - C++, F90, etc.

Pros y Contras

- Cotas exactas y estrictas
 - Quizá inalcanzables...
 - *Estricto* es quizá requerido en algunos casos
- Eficaz en tiempo de CPU
- No requiere normalidad (ignora PDF)
- No requiere diferenciabilidad
- Problemas:
 - No provee PDF del intervalo
 - No maneja correlación espacial



Método de Monte Carlo

- Monte Carlo \leftrightarrow azar (!)
- Enfoque estadístico, no determinístico
- Idea: repita para $k=1, N$
 - Generar realizaciones A_i , $i=1, m$
 - Calcule y guarde $U_k = g(A_i)$
- Luego procese los U_k generados, calculando media, varianza, etc.
- La gracia es que $\text{var}(U_k) \sim 1/\sqrt{N}$

Detalles...

- ¿Cómo generar realizaciones?
 - Asumir independencia espacial
 - Normal, media μ y varianza σ
 - demasiado fácil... y no realista
 - Modelar correlación espacial
 - No es simple; normalmente ¡hay que adivinarla!
 - Error reportado como RMS, percentil 90, etc.
 - Nada de localización espacial
 - Krigeado: simulación condicional
- Nada de esto es trivial...

Más detalles...

- Método de MC es *CPU intensive*
 - Hoy día hay CPU... y antes no
 - La CPU no es el mayor problema
- La función $g(\cdot)$ no se aproxima; se la usa directamente
- La distribución de U_k se estima mejor
- MC puede mejorarse con *bootstrapping*

El proceso requeriría...

- ✓ Evaluar la sensibilidad del modelo
- Localizar errores groseros (outliers)
- Asignar valores apropiados para los outliers y/o los faltantes

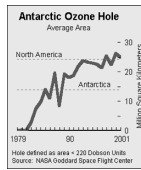


¿Qué es un outlier?

- Hay varias definiciones algo ambiguas
- *Un outlier es un valor que produce resultados inusuales (de baja probabilidad) al aplicarle cierto modelo conceptual*
 - Ej.: test de normalidad
- Suele traducirse como dato *aberrante*
- No requiere la existencia de un *valor verdadero*

¿Detección automática de Outliers?

- La Historia del agujero de Ozono
- En 1985 Farman, Gardiner y Shanklin estaban confundidos al analizar registros tomados por la misión Británica en la Antártida mostrando que los niveles de ozono habían bajado 10%.
- ¿Porqué el satélite Nimbus 7, equipado con instrumentos específicos para registrar niveles de ozono no había registrado ese descenso tan pronunciado?
- ¿Las concentraciones de ozono registradas por el satélite eran tan bajas que fueron tratadas como outliers y descartadas por un programa?



Algunos detalles...

- ¿Quién dice que es un outlier?
- En ocasiones no está claro
 - Dicotómico (ej.: digitado desde papel)
 - [mal, quizá mal, no sé, quizá bien, bien]
 - Lógica borrosa (*fuzzy*)
- Literatura estadística
 - Conjuntos pequeños
 - Errores sintéticos
 - Cálculos pesados

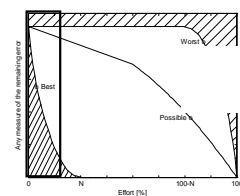
Más detalles...

- ¿Qué método usar para detectar?
 - Requiere definir relación "*mejor que*"
 - Podría automatizarse
- Casos analizados
 - Dicotómicos
 - Inspector "*perfecto*"

Tipos de errores

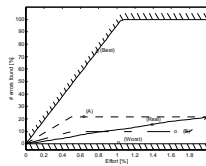
- Error Tipo I: Dato clasificado como erróneo siendo correcto
- Error Tipo II: Dato clasificado como correcto siendo erróneo
- Ventajas: *el tamaño no importa*
- Desventaja: *el tamaño podría importar*
- Se necesitarán otros estimadores

El proceso de detección



Sólo un "poco" por ciento

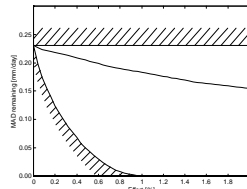
¿Cómo comparar métodos?



$$e_i = 1 - \frac{df}{dx} \frac{N}{100}$$

$$e_{ii} = \left(\frac{100-f}{100} \right) \frac{N}{100}$$

Cuando el tamaño importa...



¿Cómo sería esto *automático*?

- Datos: tipología de datos, valores, un generador de errores y N métodos
 - repetir muchas veces
 - Generar simulación de errores y contaminar banco
 - Para $i=1, N$
 - Aplicar método i
 - Contabilizar estadísticos de éxito
 - hasta lograr estabilidad estadística
 - Elegir el método con mejor resultado
- Implementado en Matlab
- Quizá costoso (pero implementado en GIS)

Un segundo problema...

- ¿Qué hacer con los datos erróneos?
 - Digitar de nuevo
 - Ir al campo a observar nuevamente
 - Resignarse
 - Eliminarlos
 - Sustituirlos
- ¿Qué hacer con los faltantes?
 - Ignorarlos
 - Sustituirlos

¡Mismo problema
¡Misma solución!

¿Cómo sería esto *automático*?

- Datos: tipología de datos, valores, un generador de huecos y N métodos
 - repetir muchas veces
 - Generar simulación de huecos y modificar el banco
 - Para $i=1, N$
 - Aplicar método i
 - Contabilizar estadísticos de éxito
 - hasta lograr estabilidad estadística
 - Elegir el método con mejor resultado
- Quizá costoso (CPU)
- Implementado (pero implementado en GIS)

Imputar es más simple...

- Imputar es un problema más clásico
 - Interpolación
 - Vecino más cercano
 - Etc.
- Varias funciones ya disponibles en GIS
- Sólo hay que simular ausencias
 - ¡No es trivial!
 - ¿Al azar? ¿en rachas?, etc.

¿Cómo comparar métodos?

- Midiendo discrepancia contra el valor conocido
- Hay un Método Óptimo
- No hay un Peor Método
- Se usan sustitutos *Naïve*
 - "El dato de ayer"
 - "El dato de al lado" "El más próximo"
 - El promedio espacial, la moda, etc.
- Tema recurrente en la literatura
- Probablemente siga siéndolo...

Plan

- ✓ Introducción
- Revisión de herramientas estadísticas
- Detectando problemas
- Imputando valores ausentes
- Ejemplos



Herramientas estadísticas

- Seguro que ya las conocen
- Necesario refrescar un poco
- Al menos algo...☺
- Univariada
- Multivariada
- Componentes Principales
- Y además...
 - Redes Neuronales
 - Krigeado

Algo básico...

- La función de distribución $F(x)$ de una variable aleatoria X se define como:

$$F(x) = \text{PROB}(X \leq x)$$

- X se dice **discreta** si

$$\text{PROB}(X = x_i) = a_i \geq 0 \forall i \text{ y además } \sum_i a_i = 1$$

- X se dice **continua** si $\text{PROB}(X=x)=0$

- La función de densidad de probabilidad $f(t)$ está definida por

$$F(x) = \int_{-\infty}^x f(t) dt$$

Esperanza matemática...

- Se define como:

► **Caso discreto** $\mu = E(x) = \sum x_i \text{PROB}(X = x_i)$

► **Caso continuo** $\mu = E(x) = \int x f(x) dx$

- También llamada **media**

- Valor modal o moda:** $x | f(x)$ es máxima

- Mediana:** $x | F(x)=0.5$

- Percentil p :** $x | F(x)=p$

- Varianza** $\sigma^2 = E((X - \mu)^2) = \int (t - \mu)^2 f(t) dt$

Exactos, pero desconocidos

El amigo Gauss...

- La distribución Normal $N(\mu, \sigma^2)$

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(t - \mu)^2 / (2\sigma^2)\right] \quad t \in (-\infty, +\infty)$$

- La "versión estándar" es $N(0,1)$

- Teo. Central del Límite

$$y_1, y_2, \dots, y_n \quad \mu_i = E(y_i) \quad \sigma_i^2 = E((y_i - \mu_i)^2)$$

$$Y = y_1 + y_2 + \dots + y_n \quad z = \frac{Y - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \quad n \rightarrow +\infty \quad \sim N(0,1)$$

Ilustración del Teo Central



- 5000 números al azar de una distribución uniforme en $[0,1]$.

► Media = 1/2, Varianza = 1/12

- 5000 números, cada uno la suma de 2 números al azar, i.e. $X = x_1 + x_2$.

► Media = 1

► Forma triangular



- Ídem, para 3 números, $X = x_1 + x_2 + x_3$

- Ídem para 12 números



Caso típico 1: datos iid.

i.e. $\mu_i = \mu, \sigma_i = \sigma$ para todo i

Teorema Central del Límite

$$\langle X \rangle = \sum_i \mu_i = N\mu \Rightarrow \langle \bar{x} \rangle = \frac{X}{N} = \mu$$

$$V(\bar{x}) = \sum_i V_i(\bar{x}) = \frac{1}{N^2} \sum_i V_i(X) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}} \quad \leftarrow \text{Famosa ley de la raíz (N)}$$

Caso típico 2: igual media

i.e. $\mu_i = \mu$ para todo i

$$\frac{\sum_i \mu_i}{\sum_i \sigma_i^2}$$

Promedio ponderado

$$V(\bar{x}) = \frac{\sum_i \mu_i^2}{\sum_i \sigma_i^2} - \left(\frac{\sum_i \mu_i}{\sum_i \sigma_i^2} \right)^2 = \frac{\sum_i \mu_i^2}{\sum_i \sigma_i^2} - \left(\frac{\sum_i \mu_i}{\sum_i \sigma_i^2} \right)^2$$

Fórmula del 'inverso de suma de inversos' para la varianza

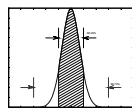
La aplicación práctica...

- Dado el banco

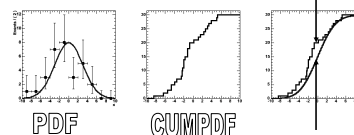
- Confirmar que es normal
 - Varios tests disponibles
 - Ej.: Test de Kolmogorov-Smirnov
- Estimar la media
- Estimar la varianza

- Dado un dato

- Calcular anomalía $|x - \bar{x}|/s$
- Comparar contra tabla



Kolmogorov-Smirnov



$$\text{Bondad de ajuste: } d = \sqrt{N} \cdot \max\{|\text{cum}(x) - \text{cum}(p)|\}$$

¡No funciona para multivariado!

Más formalmente... Test de Grubbs

- Asumiendo datos normales

- Detectar un outlier por vez, removerlo, y repetir

► H_0 : No hay outliers en los datos

► H_A : Hay al menos un outlier

- Estadístico de Grubbs

$$G = \frac{\max_i |X_i - \bar{X}|}{s}$$

- Rechazar H_0 si:

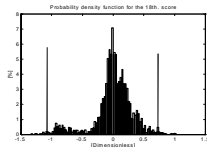
$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(N-1, N-2)}}{N-2+t^2_{(N-1, N-2)}}}$$

En general...

- Estimar los percentiles p y $1-p$
Criterio de López (¡¡!):
- Si x está en $[p, 1-p]$ → correcto
- Si no, x es outlier

También
Rousseeuw (1991)

- Habría que considerar n ,
casos
multimodales, etc.

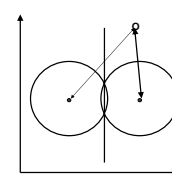


Multivariada...

- Ahora es con vectores \underline{x} ...
- $\underline{\mu} = E(\underline{x})$ es un vector
- σ^2 es ahora una matriz de covarianza C
- Anomalía era el escalar $(1/\sigma^2) * (x - \mu)^2$
- Ahora será $d^2 = (\underline{x} - \underline{\mu})^T C^{-1} * (\underline{x} - \underline{\mu})$,
también escalar
- Se denomina *Distancia de Mahalanobis*

¿Por qué tan complicado?

Caso isotrópico

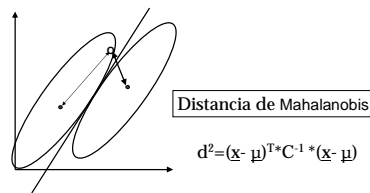


Distribuciones Gaussianas
Isotrópicas (igual varianza)

Minima distancia
cuadrada

Fuente: Mahesan Niranjan

La distancia euclídea no siempre...



Distancia de Mahalanobis

$$d^2 = (\underline{x} - \underline{\mu})^T C^{-1} * (\underline{x} - \underline{\mu})$$

Análisis de Componentes Principales

- Técnica corriente y popular
- Dada una tabla de m filas y n columnas, se "comprime" en otra de m filas y p columnas, $p < n$ y en muchos casos $p \ll n$
- Compresión *con pérdida*
- Se usa para reducir dimensionalidad del problema, conservando lo esencial de la varianza
- Imágenes multispectrales
- Datos meteorológicos

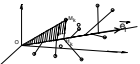
ACP(2)

- Ilustración en R^3 para $M_k - O$
- Busco e_1 tal que

$$\sum_{i=1}^n M_i H_i^2$$

sea mínima

- Luego se repite en R^2 con $(M_k - H_k)$, encontrándose e_2
- En general hay n direcciones, ortogonales entre sí



ACP(3)

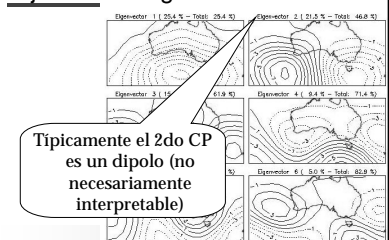
- Las proyecciones OH_k se denominan *scores*
- Hay n scores por cada fila de la tabla
- La gracia está en que...
- Se demuestra que los e_i son los Vectores Propios de C , matriz de Covarianza
- Los Valores Propios son proporcionales a la varianza de los scores
- VP pequeños ↔ scores pequeños ↔ se desprecian
- Las series de los *scores* son no-correlacionadas

ACP(4)

- Las Componentes Principales son los e_i
- También conocidas como *Empirical Orthogonal Functions* (EOF)
- Ampliamente utilizadas en Ciencias de la Tierra
- Suelen tener interpretación individual
- Pero tienen algunos problemitas...

Ej. Meteorológico

Fuente: Dr. Bertrand Timbal

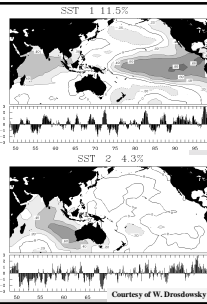


Típicamente el 2do CP
es un dipolo (no
necesariamente
interpretable)

Rotación de CP

- Facilitate physical interpretation
- Review by Richman (1986) and by Jolliffe (1989, 2002)
- New set of variable: RPCs
- Varimax is a very classic rotation technique (many others)

First two rotated PCAs of Indian/Pacific SSTs using data from Jan 1949 to Dec 1991.



Algo más que sólo Estadística...

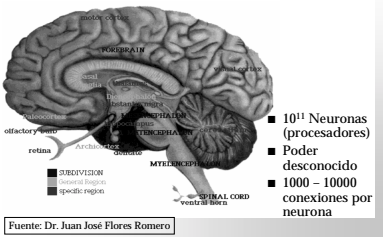
- Presentaremos ahora algo de Redes Neuronales Artificiales
- Será mencionado en Detección de outliers y en Regresión



¿Qué es una red neuronal?

- Es un modelo matemático que tiene un vago parecido con las neuronas biológicas
- La neurona es la unidad básica. En ella se distinguen las conexiones sinápticas, las dendritas (muchas), y el axón (único)
- Muchas neuronas fuertemente conectadas forman una red

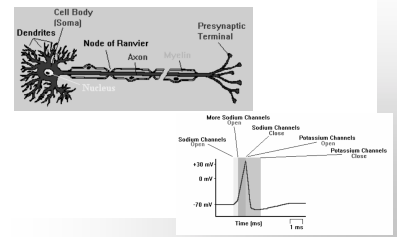
Cerebro humano...



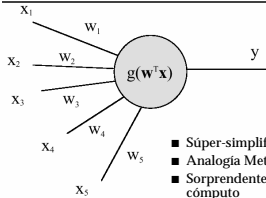
¿Cómo funciona una neurona?

- A través de las (muchas) dendritas llega la información al núcleo de la neurona
- Estimulado por esta información, se produce un efecto transmitido vía el axón
- Las conexiones sinápticas vinculan al axón con otras dendritas de otras neuronas, formando así la red

Neuronas biológicas...



Neuronas artificiales...

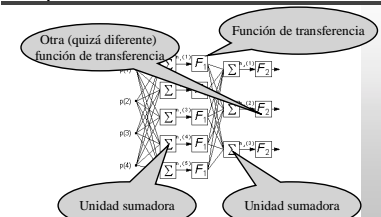


- Súper-simplificación
- Analogía Metafórica
- Sorprendente poder de cómputo

Las redes neuronales artificiales

- Simulan muy crudamente sólo algunos aspectos sustanciales de las biológicas
- La topología se modela satisfactoriamente
- Las conexiones sinápticas se modelan con coeficientes de ponderación
- La relación causa-efecto del núcleo es simulada con una función cualquiera

Esquema de una red neuronal

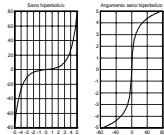


Función de transferencia (ejs.)

- Logsig

$$\text{output}_j = \frac{1}{1 + \exp\left[-\sum_i a_{ij} * \text{input}_i\right]}$$

- Senh



Algunos aspectos interesantes

- Dependiendo de la aplicación, se eligen diferentes arquitecturas de redes
- Las redes pueden utilizarse para predecir un número (output continuo), identificar una letra (output categorizado), etc.
- Toda red requiere de un “entrenamiento”
- Si la función de transferencia es no lineal, la red también lo será

Entrenamiento...

- Función objetivo (caso de regresión):

$$EMC(w) = \frac{1}{N} \sum_{i=0}^N (a_i - RN_i(w))^2$$

- Son los conocidos mínimos cuadrados (no lineales...)

Algunos términos...

- Aprendizaje
- Algoritmo de entrenamiento
- *Training set/Test set*
- Generalización
- *Overfitting*

Curso intensivo de Krigeado

- Es un método de Interpolación
- Lo hemos citado y lo citaremos en:
 - Imputación de ausencias (obvio...)
 - Detección de errores
 - Estimación de sensibilidad de modelos
- Base estadística
- Incorporado en algunos GIS (¿malamente? ¿parcialmente?...)

Curso intensivo de Krigado

- Es un método de Interpolación
- Lo hemos citado y lo citaremos en:
 - Imputación de ausencias (obvio...)
 - Detección de errores
 - Estimación de sensibilidad de modelos
- Base estadística
- Incorporado en algunos GIS
 - (¿malamente? ¿parcialmente?...)

¿Qué es la Geoestadística?

- Def.: Aplicación de la teoría de las variables regionalizadas a la estimación de procesos en el espacio
- Si $z(x)$ es el valor de z en el punto x , $z(x)$ es una variable regionalizada
 - Concepto no probabilístico
 - Quizá función continua
- Usualmente $z(x)$ está compuesta de
 - Componentes aleatorios y
 - Componentes estructurados
 - No luce "suave"
- ➔ Conviene considerar a $z(x)$ como una función aleatoria

Algunas consecuencias...

- La realidad es simplemente una realización o instancia de un experimento aleatorio
- Sólo tenemos una realidad; hay que hacer inferencia estadística sólo con ello
 - En general no sería posible
 - Requerirá hipótesis adicionales
 - Ej.: homogeneidad espacial
- Las funciones aleatorias son sólo un modelo posible de la realidad

Definiciones...

Momentos de la distribución

- 1er. orden: Esperanza $E(Z(x))=m(x)$
 - $m(x)$ es llamada "deriva" o "tendencia"
- 2do. orden:
 - Varianza $Var(Z(x))=E([Z(x)-m(x)]^2)$
 - Covarianza $C(x_i, x_j)$
 - $C(x_i, x_j) = E([Z(x_i)-m(x_i)][Z(x_j)-m(x_j)])$
 - Semivariograma $\gamma(x_i, x_j)$
 - $\gamma(x_i, x_j) = 0.5 \cdot E([Z(x_i)-Z(x_j)]^2)$
- $Var(Z(x)) \geq 0$; $\gamma(x_i, x_j) \geq 0$ pero $C(x_i, x_j)$ no se sabe

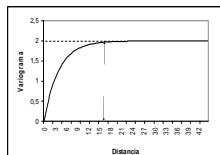
Más definiciones

- Def.: $Z(x)$ estacionaria de segundo orden si
 - $E(Z(x))$ existe y no depende de x
 - $C(x+h, x)=C(h)$ (sólo depende de la separación)
- Implica
 - $Var(Z(x))=C(0)$
 - $\gamma(x+h, x)=\gamma(h)=0.5 \cdot E([Z(x+h)-Z(x)]^2)$
- h es en general un vector; suele asumirse isotropía, por lo que $\gamma(h)=\gamma(|h|)$
- $\gamma(h)=var(Z)-C(h)$ sólo si la media es estrictamente constante; en otro caso, usar $\gamma(h)$ es más conveniente que usar $C(h)$

Más sobre variogramas...

Def.: $\gamma(h)=0.5 \cdot E([Z(x+h)-Z(x)]^2)$

- $\gamma(0)=0$; $\gamma(h) \geq 0$



Rango (Range):
Distancia a la cual el variograma se estabiliza

Meseta (Sill):
Valor constante que toma el variograma en distancias mayores al rango

Fórmula del Variograma

- El variograma debe cumplir algunas condiciones matemáticas restrictivas
- Salen de imponer que $Var(Y) \geq 0$, siendo $Y = \sum \lambda_i Z(x_i)$, λ_i y x_i conjunto arbitrario
- Hay algunos modelos de variogramas que se ajustan a los datos
 - Esférico, Exponencial, Gaussiano, Pepita, etc.
 - Hay otros menos populares
- Todos dependen de la meseta S y el rango a , excepto el denominado Pepita (nugget)

Estimación del Variograma

- Un tópico en sí mismo
- "Left to the user..."
- Métodos:
 - A sentimiento (!)
 - Mínimos cuadrados
 - Jackknife
 - Máxima Verosimilitud
 - Validación Cruzada
 - Validación Cruzada de Máxima Verosimilitud
 - ...
- Sin variograma...



Krigado

- Del geólogo sudafricano D. G. Krige
- Hay muchas variantes y casos particulares
- Caso Puntual: se modela el estimador con

$$Z^*(x) = \sum_{i=1}^N \lambda_i(x) Z(x_i)$$

eligiéndose los pesos $\lambda_i(x)$ para que sea insesgado

$$E(Z^*(x)) = m = E(Z(x))$$

y de varianza mínima $var[Z(x) - Z^*(x)]$

Algunos detalles

- Se asume m constante; hay variantes para otro caso
- Los pesos son función del punto

■ Salen del sistema:

$$\begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & 0 & \gamma_{23} & \cdots & \gamma_{2n} & 1 \\ \gamma_{31} & \gamma_{32} & 0 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \gamma_{n3} & \cdots & 0 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \\ \mu \end{bmatrix}$$

Algunos detalles⁽²⁾

- Donde $\gamma_{ij} = \gamma(x_i - x_j)$; $\gamma_i = \gamma(x_i - x)$
- Nótese que:
 - El variograma depende de los datos
 - Los coeficientes λ_i dependen del variograma, pero no de los datos mismos
 - Idem con la varianza, mediante la expresión

$$\text{Var}(Z^* - Z) = \sum_i \lambda_i \gamma(x_i - x) + \mu$$
 - La matriz del sistema es constante; puede usarse LU
- El resultado es perfectamente determinista; lo estocástico reside en los datos mismos

Algunos detalles⁽³⁾

- El de Krigado es un estimador BLUE
 - Sólo si el variograma es "exacto"
 - Sólo si la función aleatoria es normal
 - En ese caso, es el *Best* incluso comparando con los no lineales
 - Difícil de verificar la normalidad en la práctica (por lo multivariado...)
- El Krigado es interpolante
 - Sólo si se asumen datos sin error
- Bajo ciertas hipótesis
 - error $\sim N(0, \sigma^2)$; N nro. de puntos y d dimensión del espacio (típicamente 2)
 - ¡Incluso con el variograma erróneo!
 - Pero en este caso la varianza no es consistente



Algunos detalles⁽³⁾

- Si los datos *tienen* un error cuya varianza es ε^2 el sistema cambia levemente

$$\begin{bmatrix} \varepsilon^2 & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \varepsilon^2 & \gamma_{23} & \cdots & \gamma_{2n} & 1 \\ \gamma_{31} & \gamma_{32} & \varepsilon^2 & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \gamma_{n3} & \cdots & \varepsilon^2 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \\ \mu \end{bmatrix}$$

Simulación

- Def.:
 - No condicionada: Consiste en generar realizaciones con igual media y varianza que la disponible
 - Condicionada: Idem, pero obligando a que además adopte valores específicos en ciertos puntos
- Tres tipos de métodos
 - Espectrales, Bandas Rotantes y Matricial
 - Sólo presentaremos el Matricial



Método Matricial de Simulación

- No es el más eficaz si se necesitan muchos puntos
 - Matricial $O(n^3)$
 - Bandas rotantes $O(n^{1/2})$
- Implícitamente se asume normalidad
- La fórmula para Z_S es:

$$Z_S = Z^* + \mathbf{M} \cdot \mathbf{u}; \mathbf{M} \cdot \mathbf{M}^T = \mathbf{C}; \mathbf{u}_i \sim N(0, 1)$$
- La simulación se logra generando diversos \mathbf{u}
 - Problema estándar
 - Muchas librerías disponibles

¿Para qué se usa la Simulación?

- Generar realizaciones
 - Compatibles con las medidas disponibles
 - Compatibles con el variograma asumido
- Ejemplo: un MDE
 - Generar N realizaciones del raster buscado
 - Delinear zona de visibilidad a un mástil
 - Calcular área A_i de esa zona
 - Calcular valor esperado, promedio, máximos, etc. del conjunto A_i y sus niveles de confianza
- Se comentarán más casos luego

Literatura & Software

- Digital:
 - Rudolf Dutter, Vienna Inst. of Technology
 - CD del curso: http://www.statistik.tuwien.ac.at/public/datt/vorles/geom_05/geom.html
 - Denis Marcotte, École Polytechnique de Montréal
 - CD del curso: <http://geom.polytechnique.ca/~marcotte/gls400/geom.html>
 - Oscar Rondón, Venezuela
- Papel:
 - Samper, F.J. y Carrera, J. 1990. Geostatística: Aplicaciones a la hidrología subterránea. CIMNE, ISBN 84-404-6045-7
- Biblioteca GSLIB
- Matlab+EasyKrig
 - http://globec.whoi.edu/pub/software/kriging/V2.1/easy_krig2.1

Módulo 4: Mejorando la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org

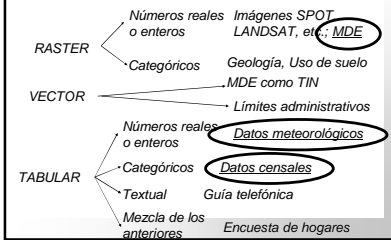
Parte II

Plan

- ✓ Introducción
- ✓ Revisión de herramientas estadísticas
 - Detectando problemas
 - Imputando valores ausentes
 - Ejemplos



Una posible categorización de datos



El cómo de la detección de outliers

- Métodos tradicionales para el caso multivariado
 - Distancia de Mahalanobis

$$\frac{(x-\bar{x})}{\sigma} \leq \Rightarrow (x-\bar{x})^T C^{-1} (x-\bar{x})$$
 ¿Cómo hallar C y \bar{x} ? => Clásico, MCD, MVE, Hadi (1994), Roche (1996), etc.
 - Análisis de Componentes Principales (PCA)
 - Hawkins, 1974; López, 1994a,b, 1996, 1997
 - Otros métodos...

Mahalanobis de vuelta...

- Si $d^2(x) = (x-\bar{T})^T C^{-1} (x-\bar{T}) > d_{crit}$ → outlier
- Depende de cómo se construyen C y \bar{T} puede ser inapropiado si hay outliers (!)
- Ej: Philips data

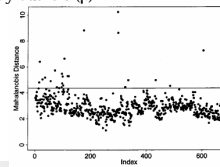
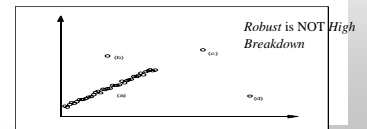


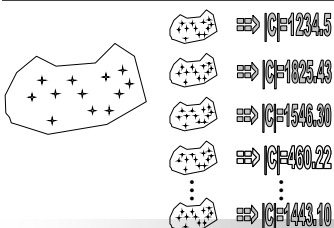
Figure 1. Plot of Mahalanobis Distances for the Philips Data.

¿Cómo hallar C y \bar{T} ?

- Estimadores clásicos de media y varianza
- Estimadores robustos
- High Breakdown estimators



MCD (Rousseeuw *et al.*, 1987)



FAST-MCD (Rousseeuw *et al.*, 1999)

- Mismo criterio, otro algoritmo
- Más rápido, etc.
- Maneja "exact fit"

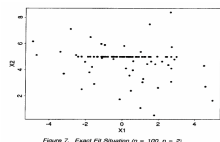


Figure 7. Exact Fit Situation ($n = 100, p = 2$).

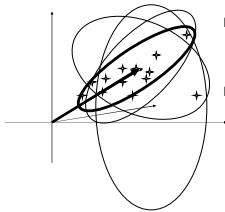
Hadi (1992,1994)

- Similar al MCD, pero no combinatorio
- Más rápido que otros
- Implementado en SAS y otros paquetes estadísticos

¿¿¿Y la esfinge???



MVE (Rousseeuw *et al.*, 1990)



- C y T tal que hay ε datos fuera de cada elipse
- C y T no son ahora función directa de un subconjunto de datos

Comentario...

- Los anteriores son casos particulares de estimadores más generales
 - Estimador-S
 - Estimador-M
- Veamos apenas una definición de cada uno de ellos

Estimador-S

- $C=C(X)$ y $T=T(X)$ tales que:

$$\begin{cases} \det(C) \text{ sea mínimo} \\ \frac{1}{n} \sum_i \rho(d_i) = b_0 \end{cases}$$

$$d_i = \sqrt{(x_i - T)^T C^{-1} (x_i - T)}$$

- $\rho(d)$ función no decreciente
- MVE: ρ pertenece al conjunto $\{0,1\}$

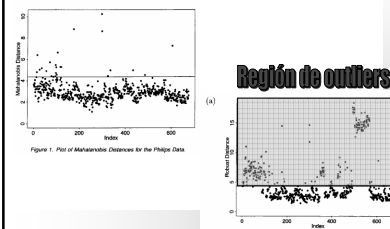
Estimador-M

- $C=C(X)$ y $T=T(X)$ tales que:

$$\begin{cases} \det(C) \text{ sea mínimo} \\ \sum_i (x_i - T)^T u_1(d_i) = 0 \\ \frac{1}{n} \sum_i (x_i - T)(x_i - T)^T u_2(d_i^2) = C \end{cases}$$

- $u_1(d)$ y $u_2(d)$ ni negativas ni decrecientes si $d > 0$

Aplicación: Philips data



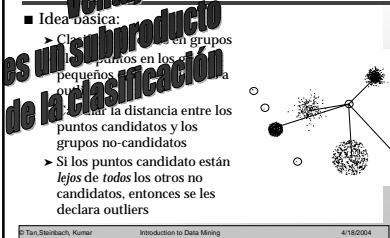
Recapitulando...

- Estimadores basados en Mahalanobis
- Difieren entre sí en la forma de estimar unos C y T *apropiados*
- Datos C y T son *simples* de aplicar
- Matemáticamente tratables
- Para datos tabulares
- Requieren datos sin ausencias
 - ➔ quizá hay que imputar primero

Minor ACP (Hawkins 1974)

- Los CP *mayores* (i.e. con gran valor propio) tienen interpretación física
- Normalmente se retienen, y los *menores* se descartan
- Hawkins propone utilizar los *scores* asociados como detectores de errores
- Son típicamente pequeños, e indican algo inusual cuando son grandes

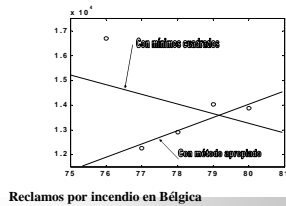
Otras ideas: Conglomerados



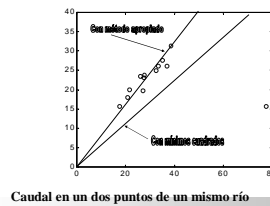
Detección en el contexto de regresión

- Idea: ajustar con una función, analizar las discrepancias y separar las más groseras
- Problema: los errores pueden enmascarse unos a otros
- Problema: los errores pueden afectar significativamente la función de ajuste (Ejemplo: OLS)
- Solución ==> High breakdown methods (LTS, LMS, etc.)

Efecto de errores en regresión⁽¹⁾



Efecto de errores en regresión⁽²⁾



OLS (Gauss, 18XX?)

- Minimiza la *suma* de cuadrados de residuos
- Sensible a outliers en varias formas
- Muy afectado por *enmascaramiento*
- ¡Implementado everywhere!
- En problemas tabulares tolera ausencias
 - Requiere un OLS por cada combinación de ausencia/presencia ➔ puede ser pesado...

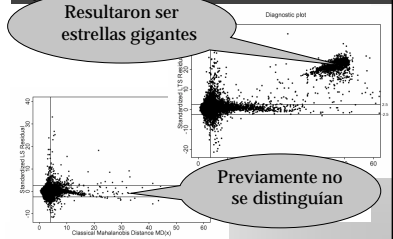
LMS (Rousseeuw 1984)

- Minimiza la *mediana* de los residuos
- Es como OLS si se descartara *cierto* 50% de la población
- Es por lo tanto del tipo *High Breakdown*
- Problema combinatorio ➔ ¡explota!
- Existen alternativas no combinatorias (Hawkins 1993)
- Igual que OLS: tolera ausencias

LTS (Rousseeuw 1984)

- Minimiza la *suma ponderada* de los residuos
- Pesos se eligen del conjunto {0.0, 1.0}
- El total de casos 0.0 se especifica a priori
- Es por lo tanto del tipo *High Breakdown*
- Problema combinatorio ➔ ¡explota!
- Existen alternativas no combinatorias (Hawkins 1993)
- También tolera ausencias

Ejemplo: Rousseeuw et al., 1999



Datos raster: caso del MDE

- Ampliamente estudiado en Agrimensura
- Antes: pocos puntos, muy precisos
 - Típicamente formato TIN
 - Raster se *calculaba* a partir del TIN
- Antes: fotogrametría aérea
 - Típicamente curvas de nivel
 - Raster se *calculaba* a partir de las curvas
- El productor dispone de controles internos
- También se detectan errores al comparar con la hidrografía, etc.

Situación presente

- Surgen otros métodos de creación (satélite, GPS, LIDAR, etc.)
- Muchos más puntos, algo menos precisos
 - Imagen de satélite, etc.
 - Se genera directamente el raster
- ¡El usuario puede ahora ir al campo y controlar!
- Detalle esencial: un pixel puede estar errado sin implicar al vecino
- Amerita otros métodos de control...



Método de Hannah (1981)

- MDE en formato raster
- Establece límites en la pendiente y cambio de pendiente
 - ¡Requiere especificarlos a priori!
- Usa interpolante local
- Fácilmente implementable en GIS
- Poco impacto en la literatura; mencionado aquí por *completness*

Método de Felicísimo (1994)

- Imputa interpolando con los vecinos
- Cualquier interpolante sirve; propone polinomio de 2do. grado en (i,j)
- Analiza la distribución de la diferencia del interpolado vs. el verdadero valor
- Asume normalidad, y saca límites
- Relativamente simple, implementable en SIG
- Veremos un ejemplo más adelante

Métodos mixtos

- Usan indirectamente métodos de regresión para detectar los errores
 - Uso de la verosimilitud (*likelihood*)
 - Interpretación de los roles de las neuronas en redes neuronales artificiales



Función de Verosimilitud (López, 1997)

- En un contexto de Kriging aparece el Variograma
- Depende de: Tipo, Alcance α y Meseta S
- Método de VCMV (Samper *et al.*, 1987)
 - Elegir α y S que maximicen la VCMV asumiendo que no dependen del tiempo...
- ¿Cómo es el método de VCMV?

VCMV (Samper, 1987)

- VCMV: Elijo α y S , y para cada fecha repito para los n puntos disponibles:
 - Retiro el i -ésimo
 - Interpolo mediante kriging
 - Conservo la discrepancia observada
- Luego se calcula la Verosimilitud
- Nuestro n era relativamente pequeño
- En general el proceso requiere minimizar una función no lineal costosa...

Nuestro problema particular de VCMV

- En realidad, nosotros no necesitábamos α y S ...
- Sólo interpoláramos en los puntos dato
- Para nuestros fines sólo necesitábamos una C y T obtenida de los datos experimentales
 - No hubo necesidad de minimizar la función
- Se asumió homogeneidad e isotropía
- Se asumió también α y S constantes en el tiempo
- Idea: dados α y S , la Verosimilitud pasa a ser función del tiempo; un número por día
- Días con valores extremos → outliers!

Reflexión...

- Casi nadie usa o referencia la Verosimilitud misma; sólo la maximiza
- Algo parecido a los Mínimos Cuadrados
 - ¿alguien se fija si los mínimos cuadrados son pequeños?
- Resultó ser uno de los mejores métodos en nuestros experimentos

No por trillado el camino es conocido
López (2005)

Métodos mixtos

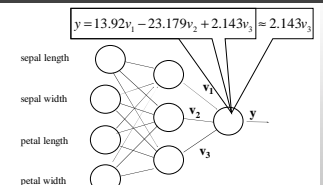
- Usan indirectamente métodos de regresión para detectar los errores
 - Uso de la verosimilitud (*likelihood*)
 - Interpretación de los roles de las neuronas en redes neuronales artificiales

Uso de redes neuronales

- Se reconocen dos líneas posibles
- Línea 1: Clasificación
 - Clasificar en forma no supervisada
 - Clusters con pocos elementos → outliers
 - Línea 2: Regresión
 - Ajustar por MC y analizar discrepancias
 - Línea 2.5: Regresión+...
 - Ídem 2, pero luego interpretar roles
 - Unpublished work, by López

ANN para regresión

Tomado de Benítez *et al.*, 1997



Proponen simplificar la red...

ANN para regresión

y la clasificación anda bien

Alta. Coeficientes grandes
= división de números

¿Qué rol tenían las otras dos?

Versión modificada

$$y = 13.92v_1 - 23.179v_2 + 2.143v_3$$

if ($v_2 > 0.45$ or $v_1 > 0.73$)
then $z = 1$
else $z = 0$

¡No participa!

Rol \iff coeficiente

Ventajas...

- La Red se entrena como siempre para regresión/clasificación
- Se inspeccionan los pesos; no hay que reentrenar
- Los outliers no se decretan; ¡surgen!
- Desventaja: los pesos pueden ser muy sensibles a los outliers \rightarrow *masking*
- Fue testeado en el ejemplo (caso pequeño, *de paper*) y con lluvia, etc.
- ¡Fue el óptimo!
- Es aún una teoría. Queda mucho por hacer...

Ejemplos de detección de outliers

- Comentaremos algunos casos
- Tabular Cuantitativo: datos meteorológicos
 - Observados en una red de puntos fijos
 - Muchas medidas en el tiempo
 - Viento horario
 - Fuerte correlación espacio-tiempo
 - Lluvia diaria
 - En Uruguay, sólo correlación espacial
- Tabular Categórico: Datos de un Censo
- Raster: MDE

Datos tabulares: lluvia y viento

- Usamos lluvia diaria y viento horario
 - Lluvia tiene sólo correlación espacial
 - Viento tiene espacio-temporal
- Para el viento, 35% de los errores simulados aparecieron en el primer paso de depuración
- Para lluvia, 81% de los errores simulados aparecieron en el primer paso de depuración

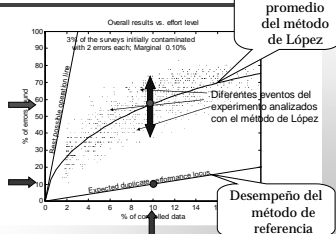


Datos tabulares: censo nacional

- Sólo para datos categóricos puros
- Pudimos remover 50% de los errores revisando un 10% del conjunto
 - Cinco veces mejor que digitar de nuevo
- Método general, automatizable, basado en ACP



Gráficamente...

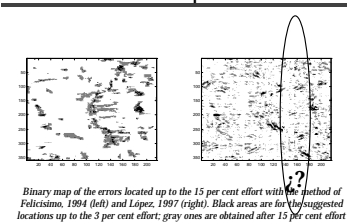


Datos raster: MDE (López 1997)

- Buscamos algunos tipos de errores
 - Salt and pepper
 - Spike
 - Pyramid
- El método es aplicable para cualquier raster cuantitativo (imágenes, fotos, etc.)
- En el artículo, 40% de los errores fueron encontrados con probabilidad $> 88\%$
- Podría ser una herramienta útil para productores y usuarios



Felicitísimo vs. López ☺



Resumen: No matar al mensajero...

- No puede des-inventarse el GPS
- No puede des-inventarse el GIS
- No pueden des-digitalizarse los mapas
- Hay que *entender* los problemas
- Hay que *tomar* decisiones apropiadas
- Hay que *aprender* de otros ejemplos

¿Estamos perdidos?



Plan

- ✓ Introducción
- ✓ Revisión de herramientas estadísticas
- ✓ Detectando problemas
- Imputando valores ausentes
- Ejemplos



Recordemos: ¿Porqué imputar?

- Detectado un error...
- Modelos que no toleran ausencias
- Bajar costo al medir menos
- Típicamente métodos de *Interpolación*
 - Medidas escasas, de alta exactitud
- Actualmente más y más *Aproximación*
 - Más automatismo, menos control humano
 - Medidas abundantes, de menor exactitud



Enfoque es función del dato...

- Datos "*puramente*" espaciales
 - Caso más familiar para la audiencia
 - Métodos de Interpolación:
 - TIN, Splines, Kriging, Cressman, etc.
 - Ej.: MDE, tipo de suelo, etc.
- Datos espacio-temporales
 - Correlación espacial + temporal
 - Ciencias de la Tierra, pero no Agrimensura
 - Ej.: Meteorológicos, uso del suelo, etc.
- Formulación sensiblemente específica

Datos *puramente* espaciales

- En la gran mayoría son Métodos lineales
- Coeficientes son función de punto
- Toleran ausencias
- A veces son lineales pero complicados
 - Cokriging
- Hay también métodos no lineales
 - Redes neuronales
 - Ecuaciones constitutivas (EDP)

Datos espacio-temporales

- Típicamente equi-muestreados en el tiempo
- Problema no resuelto: covarianza cruzada tiempo-espacio
- Muy usual en las Ciencias de la Tierra
 - Ej.: Meteorología, Hidrología, etc.
- Habitual en las aplicaciones GIS
 - Ej.: Tráfico/Tránsito, uso del suelo (!)
- Poco o mal manejo en GIS comerciales

Muchos métodos...



Procedimiento sugerido...

- Repita un número grande de veces
 - Generar ausencias al azar
 - Imputar con método1, método2, etc.
 - Calcular estadísticos de ajuste (distancias)
- Comparar estadísticos, y luego elija...
- Ventajas:
 - Tiene base estadística
 - Lo puede hacer el productor o el usuario
 - ¡No requiere ir al campo a medir!
- ¿Y las desventajas?

Desventajas o problemas...

- No todos los métodos están en los GIS
- ¿Cómo generar ausencias?
 - Al azar (MCAR)
 - En rachas (usual en datos meteorológicos)
- Hay que caracterizar primero SUS ausencias
- Otro tema: los estadísticos de éxito
 - Datos cuantitativos
 - Datos categóricos
 - Considerar o no el *impacto en el modelo*
- Un detalle más: el tiempo de cálculo

¿Cómo generar ausencias?

- Es más fácil que generar errores
- Hipótesis inicial: MCAR
 - Test descrito en Little (1988)
- En la práctica también había *rachas*
 - Rotura de instrumento
 - Pérdida de documento original en papel
- Quizá parezca excesivo detalle, pero...

Estadísticos de éxito

- Métricas usuales:
 - RMSE: Da mucho peso a errores groseros
 - MAD (Promedio): ídem RMSE
 - Percentiles: quizá más apropiado
- Asumiendo que existe un dato *verdadero* existe un Método Óptimo que lo asigna
- No existe en cambio un Peor Método
 - Podría usarse un *Naive* como referencia

Más sobre Estadísticos

- Podría considerarse el modelo
 - Errores sistemáticos pueden ser peores que errores groseros
 - Groseros son detectables; sistemáticos no
 - Ej.: errores en una factura:
 - Sesgados: ¡el cliente se queja dependiendo del signo!
- Otro problema: RMSE vs. Exactitud original
 - Ej.: RMSE lluvia ~7 mm/día; Exactitud 5 mm/día, pero ¡¡precisión 0.1 mm/día!!

Plan

- ✓ Introducción
- ✓ Revisión de herramientas estadísticas
- ✓ Detectando problemas
- ✓ Imputando valores ausentes
- Ejemplos



Caso del Viento horario

Problema:

- Completar un banco de datos de viento de superficie horario
- Comparar diferentes métodos, en dos diferentes casos:
 - Ausencias al azar
 - Ausencias planificadas

Fuente: Proyecto CONICYT/BID 51/94 (1999)

Diseño de la metodología

- Seleccionar un banco apropiado, lo más completo posible
- Ocultar temporalmente los valores a ser imputados (elegidos al azar o no)
- Para cada método
 - imputar todos los valores ausentes
 - calcular RMSE y MAD de las discrepancias entre el valor real y el imputado

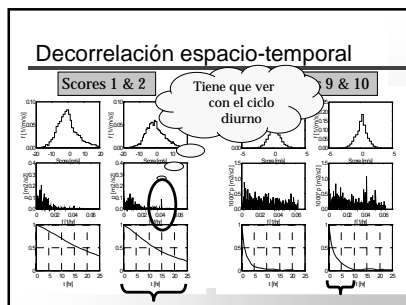
El banco de datos

- Cinco estaciones meteorológicas, separadas no más de 400 km, en terreno suave
- Aproximadamente 25 meses de registros horarios



Descripción de los métodos

- Interpolación Óptima (equivalente a Krigado Ordinario)
- Métodos basados en el Análisis de Componentes Principales:
 - Temporal Interpolation of Principal Scores (TIPS)
 - Penalty Of the Principal Scores (POPS)



Resultados preliminares obtenidos

a) Ausencias sistemáticas

- Se asumieron tres lecturas diarias (8, 14 y 21 hs.), en cuatro de cinco estaciones
- Con TIPS se logra un RMSE de 2.05 m/s
- Con POPS se logra RMSE de 2.84 m/s
- La Interpolación Óptima produce 2.84 m/s
- Asignando simplemente la media histórica el RMSE es de 3.24 m/s

Resultados preliminares obtenidos

b) Ausencias al azar

- Se ocultó aleatoriamente un 20% de los datos, criterio MCAR
- Con TIPS se logra un RMSE de 1.67 m/s
- Con POPS se logra RMSE de 2.33 m/s
- La Interpolación Óptima produce 2.37 m/s
- Asignando la media histórica el RMSE es de 2.76 m/s

Conclusiones

- El uso de la información temporal da resultados más precisos, sugiriendo un muestreo excesivo para esta zona
- Los resultados deben ser corroborados en ensayos más extensos, para darle validez estadística
- Otros métodos deben ser incluidos en la comparación

Ver informe final de 1999

Caso de la lluvia diaria

- Nuevamente, un problema *tabular*
- 10 estaciones, registros diarios (mm/día)
- Correlación espacial pero no temporal
 - TIPS falla miserablemente
- Problema difícil
 - RMSE del Mejor vs. Peor método evaluado difieren en 30%
- Mejor RMSE: 7 mm/día; según los expertos, la Exactitud~5 mm/día (!)

Sugerencias para lectura...

- Informe CONICYT/BID 51/94 (1999)
 - Análisis comparativo de ~30 métodos
 - Imputación
 - Detección de outliers
 - Descripción de métodos, referencias, etc.
 - No orientado a meteorología
 - Único estudio sistemático conocido

Módulo 4: Mejorando la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org

Módulo 5: Manteniendo la Calidad

Carlos López Vázquez

carlos.lopez@ieee.org

Plan

- Control en una Cadena de Producción
 - Introducción
 - ¿Qué medir?
 - Sistemas de Medición
 - Implementación
 - Resumen
- Conceptos de ISO 9000



Introducción

- Producción de datos vs. Bienes
- Producto único
 - Similar a un astillero
 - Producto individual
 - Herramientas, operarios, etc. alrededor del bien
 - CC por inspección, al final
 - Poco repetitivo
 - No puede hacerse estadística
 - Ej.: Construcción de carretera

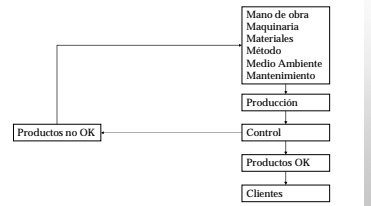


Producto en serie

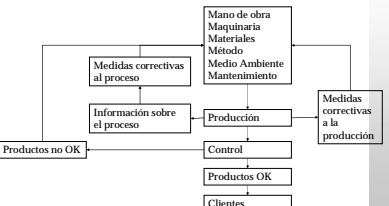
- Similar a una fábrica
 - Se mueve el bien, no los técnicos
 - Obreros especializados
 - Proceso establecido, aplicado muchas veces
 - Muy repetitivo
 - CC estadístico, sobre el proceso
- Ej.:
 - Serie cartográfica
 - Datos Meteorológicos



Control al final

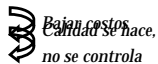


Control durante el proceso



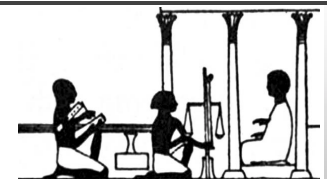
Algún concepto

- El CC ha variado en el tiempo
 - Verificación
 - CC del Producto
 - CC del Proceso



Verificación

- Se aplica desde la antigüedad
- Se realiza la inspección sobre el 100% del lote
- Se realiza la inspección sobre el trabajo/producto ya terminado
- Sólo se inspecciona el producto o el resultado del servicio (no el diseño, ni el proceso, ...)



Control de Calidad previo al almacenamiento de cosechas. De una tumba en Beni Hasan, Egipto, ca. 1900 a.C.

Verificación

- Pretende ser un filtro que proteja al consumidor
- Suele ser realizada por personas distintas a quienes realizan la producción (p.ej.: el Depto. de Calidad)
- Con ello se gana imparcialidad, pero se pierde la implicación de Producción en la calidad

Control Estadístico

- Origen en los años 20
- Formalización y extensión a partir de la Segunda Guerra
- Se basa en identificar y tratar características de Calidad como variables aleatorias



Control Estadístico del Producto

- Menor costo que Verificación
- El Control puede ser asumido por Producción
- Inconveniente: sigue actuando sobre el producto acabado
- Se aplica para Control de Aceptación (normas MIL-STD 105 y 414)

Control Estadístico del Proceso

- Foco en el proceso de producción, y no en el producto terminado
- Busca identificar Causas de la variabilidad, las que inciden (negativa o positivamente) en la Calidad
- Una vez identificadas hay que eliminarlas

Algún dato...

- Con sólo control al final se detecta menos del 80% de los errores
- Mejor ir al Proceso
- Dos problemas similares/diferentes:
 - Bien propio (producción interna)
 - Bienes de proveedores (id. externa)
- Soluciones similares/diferentes:
 - SPC
 - Muestreo

Soluciones similares...

- Requieren medir propiedades relevantes
 - Cuantitativas
 - Ej.: error menor a 5.2 mt
 - Inspección por Variables
 - Cualitativas
 - Ej.: el papel está doblado de acuerdo a la norma
 - Inspección por Atributos
- Obvio que hay casos mixtos
- Al final hay rechazo/aceptación
 - Niveles de confianza
 - Mínimo costo

Plan

- Control en una Cadena de Producción
 - ✓ Introducción
 - ¿Qué medir?
 - Sistemas de Medición
 - Implementación
 - Resumen
- Conceptos de ISO 9000



¿Qué medir?

- Hay varias maneras de medir desempeño
 - Conteo de datos erróneos
 - Reglas de Negocio no cumplidas
 - Ingreso duplicado
 - Tamaño de datos erróneos
 - Efecto Godzilla
 - Efecto de datos erróneos
 - Número de Quejas de Clientes

Reglas del Negocio

- Def: Relaciones que tienen que cumplir los datos válidos
 - Sustanciales: $Lluvia > 0 \leftrightarrow Nubosidad > 0$
 - Formales: $Lluvia\ diaria \geq 0$
- Surgen de:
 - Expertos en el tema
 - Análisis estadístico de datos



Reglas de Expertos

- Son personales
- Son incompletas
- Son de difícil actualización (¡Hay que matar al experto!)
- ¡No requieren datos!
- Suelen redundar
- Hay software específico

Análisis Estadístico

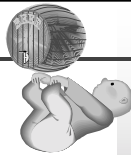
- Hoy conocido como "Minería de Datos"
- Busca relaciones predictivas empíricas
 - No verificadas \leftrightarrow ¿erróneos?
- Busca describir clases
 - No clasificados \leftrightarrow ¿erróneos?
- Registros individuales \rightarrow conclusiones globales

Modelos predictivos

- Modelos de regresión
 - Vol. Compra = $f(\text{ingreso, edad, nro. de hijos})$
- Típicamente vinculan datos cuantitativos
- Se analiza $y = | \text{Compra} - f(x) |$
 - Si $y \leq y_0 \rightarrow \text{ok}$
 - Si $y > y_0 \rightarrow \text{¡sospechoso!}$
- Para corregir se requiere acceso al "verdadero valor"

Clasificación

- Buscan patrones en los datos
 - Cerveza+pañales+pizza
 - Perfume+bombones-leña
 - Mañana de domingo+chorizos
- Vinculan datos categóricos y cuantitativos
- Si alguien compra Leña y Perfume ...
- Para corregir se requiere acceso al "verdadero valor"



Plan

- Control en una Cadena de Producción
 - ✓ Introducción
 - ✓ ¿Qué medir?
 - Sistemas de Medición
 - Implementación
 - Resumen
- Conceptos de ISO 9000



Sistemas de Medición

- Tema muy importante
- Las Métricas definen el comportamiento
- Peligros
 - Métricas inapropiadas
 - Administrar Métricas y no el proceso
- Ejemplos

El caso de la Industria siderúrgica

- Produce chapas, perfiles, varillas, etc. en serie
- Objetivo: producción bruta en Ton/mes
- Cambios -8 hs.
- Grandes stocks
- Demora en cumplir pedidos



Resultado:

Ton/mes cumplidas, y clientes insatisfechos

Administrar la métrica

- VW nombra CEO español
- Año 1: récord de ganancias
- Año 2: nuevo récord de ganancias
- Año 3: ¡CEO despedido!

- Corto plazo vs. Largo plazo



Medir vs. no Medir

- Lo esencial es invisible a los ojos
- Medir mal es peligroso



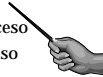
Plan

- Control en una Cadena de Producción
 - ✓ Introducción
 - ✓ ¿Qué medir?
 - ✓ Sistemas de Medición
 - Implementación
 - Producción interna
 - Producción externa
 - Resumen
- Conceptos de ISO 9000



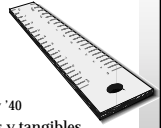
Producción interna

- Introducción
- Variabilidad del proceso
- Estabilidad del proceso
- Límites de control:
 - Teoría y métodos
 - Fórmulas
- Resumen



Introducción

- "...Si no puede medirse..."
- Medir ↔ Controlar
- SQC
 - Shewhart, 1920
 - Producción en serie
 - Ideas consolidadas en los '30 y '40
- Corriente para manufacturas y tangibles
- Aplicable para intangibles



Propósito del SQC

- Definición:
 - Predecir el comportamiento futuro del proceso
- Algunas características
 - Orientado al futuro
 - Usa datos pasados y actuales
 - No es lo mismo que auditoría
- ¿Qué tipo de predicción?
 - "...si no pasa nada nuevo, todo seguirá igual..."
 - "...busque bien, porque algo ha pasado..."



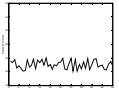
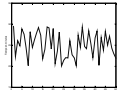
Producción interna

- ✓ Introducción
- Variabilidad del proceso
- Estabilidad del proceso
- Límites de control:
 - Teoría y métodos
 - Fórmulas
- Resumen



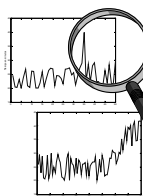
Predicciones...

- Proceso Bajo Control
- Se puede observar:
 - Una línea media
 - Un rango estable



➔ Hay previsibilidad en el rango

Predicciones...

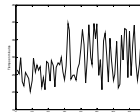


Proceso Fuera de Control

- Se puede observar:
 - Una accidente
 - Anomalías
- ➔ Hay que buscar una Causa asignable

Predicciones...

- Otro ejemplo:
- La tendencia no cambió, pero sí el rango



*Estas cosas pasan...
Hágalo Ud. Mismo*

Producción interna

- ✓ Introducción
- ✓ Variabilidad del proceso
- Estabilidad del proceso
- Límites de control:
 - Teoría y métodos
 - Fórmulas
- Resumen

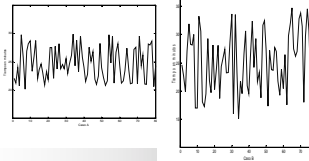


¿Cómo es un proceso estable?

- Proceso estable ↔ parámetros “estables”
- Parámetros estadísticos
 - Media
 - Mediana
 - Desviación estándar
 - Rango
 - Quintiles
 - Otros...

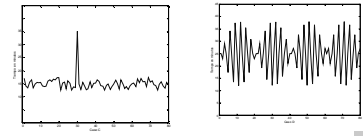
¿Cómo es un proceso estable?

Dos casos estables, con la misma media pero diferente rango



¿Cómo es un proceso estable?

Tiene la misma media, pero...

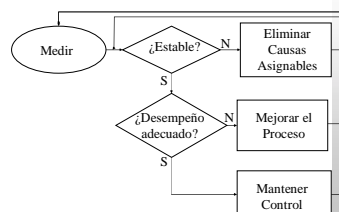


En resumen...

- Primer problema: medir
- Segundo problema: decidir si el proceso es estable (*está bajo control*) o no
- Tercer problema: qué hacer en cada caso
 - Corregir Causas Asignables
 - Mejorar el Proceso

Todo puede esquematizarse...

Un diagrama de flujo



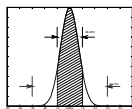
Producción interna

- ✓ Introducción
- ✓ Variabilidad del proceso
- ✓ Estabilidad del proceso
- Cartas de control:
 - Teoría y métodos
 - Fórmulas
- Resumen



Cartas de Control

- Se basa en trabajos de la década de 1920-1930
- Sólo estadística básica... ¡o todo lo compleja que se quiera!
- Recordemos algo de la distribución Normal...



Teoría

- Dados γ_t , $t=1,2,\dots,k-1$
- SI γ_t es aleatoria y SI $\gamma \sim N(\mu, \sigma)$ entonces γ_t , $t=k+1, k+2, \dots$ será aleatoria
 - “Aleatorio” → difícil de probar
 - “No Aleatorio” → criterios
 - 3σ (tres sigmas)
 - Rachas

3σ : algunas fórmulas...

- $n > 25$

- Fórmulas básicas

$$\mu = \bar{p} = \frac{1}{n} \sum_{i=1}^n p_i \quad \sigma = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$CL \approx \bar{p}$$

- Hay casos especiales, con fórmulas idem

3σ: Casos corrientes y especiales

p_i (o γ_i) pueden ser:

- una magnitud directamente asociada a una observación (ej.: tiempo de viaje)
- el promedio de una magnitud en un lote de N unidades ($N > 2$).
- el número total de fallas observado en un lote de N unidades
- la proporción de defectuosos en un lote de N unidades. Sólo hay unidades buenas o malas.

habrá fórmulas para cada caso

Control de Rachas

Alguno por encima de $CL+3\sigma$	UCL
2 de 3 por encima de $CL+2\sigma$	2σ
4 de 5 por encima de $CL+1\sigma$	1σ
8 consecutivos por encima de CL	CL
8 consecutivos por debajo de CL	1σ
4 de 5 por debajo de $CL-1\sigma$	2σ
2 de 3 por debajo de $CL-2\sigma$	3σ
Alguno por debajo de $CL-3\sigma$	LCL

Producción interna

- ✓ Introducción
- ✓ Variabilidad del proceso
- ✓ Estabilidad del proceso
- ✓ Cartas de control:
 - Teoría y métodos
 - Fórmulas
- Resumen

En resumen:

- Hay lógica detrás de SQC (y teoría)
- SQC define *Estabilidad del proceso* y da:
 - Condiciones
 - Predicciones
- SQC ayuda a *detectar* problemas
- Una vez detectados, hay que resolverlos
- También sirve para confirmar que los niveles de calidad son inalcanzables
- En ese caso hay que cambiar los procesos

Plan

- Control en una Cadena de Producción
 - ✓ Introducción
 - ✓ ¿Qué medir?
 - ✓ Sistemas de Medición
 - Implementación
 - ✓ Producción interna
 - Producción externa
 - Resumen
- Conceptos de ISO 9000

Planes de aceptación por muestreo

- Si acepto el lote sin hacer CC
 - Costo nulo del CC (obvio)
 - Costo de la No Calidad
- Verificación 100% sería lo ideal
 - Caro, lento etc.
- ¿Algo intermedio?
- Muestreo → lote=N, verifico sólo n(N), y descarto o no el lote entero

	Costo del Control de Calidad	Costo de la No Calidad
Aceptar todo	Nulo	Alto
Verificar todo	Alto	Nulo

Algo intermedio...

Activa 2002

Ventajas del muestreo

- Menor costo total
- Pueden identificarse dos riesgos:
 - Comprador: aceptar lote con muchos defectos
 - Vendedor: que le rechacen un lote bueno
- Tratamiento estadístico es ahora viable
- Criterios económicos
- Inspecciones normal, rigurosa y reducida
- Muestreo simple, doble o múltiple

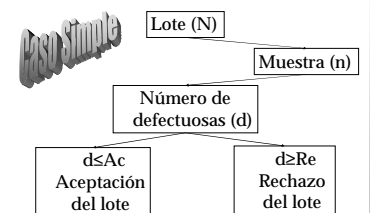
	Costo del Control de Calidad	Costo de la No Calidad
Aceptar todo	Nulo	Alto
Verificar parte	Medio	Medio
Verificar todo	Alto	Nulo

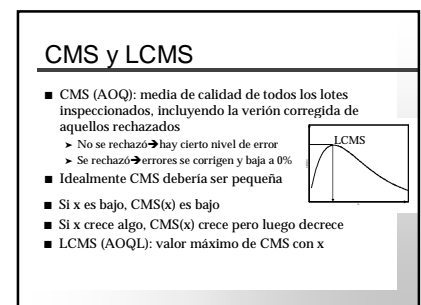
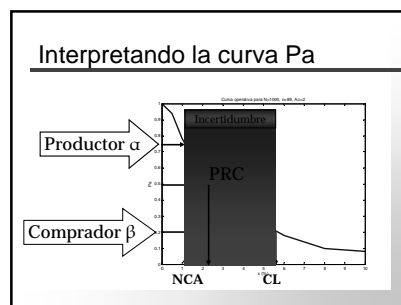
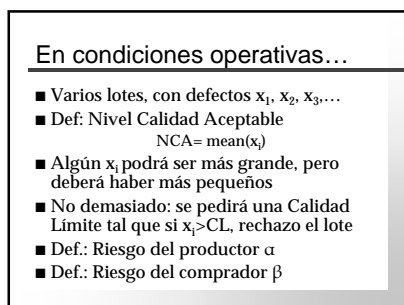
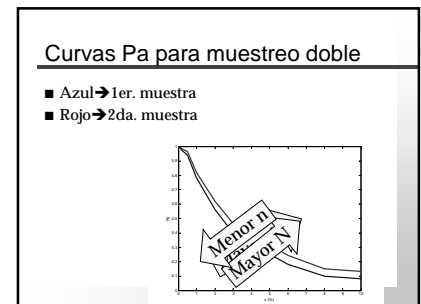
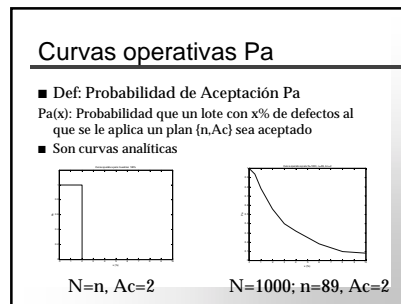
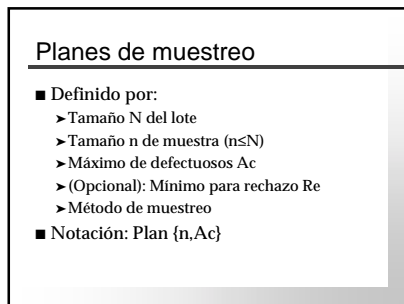
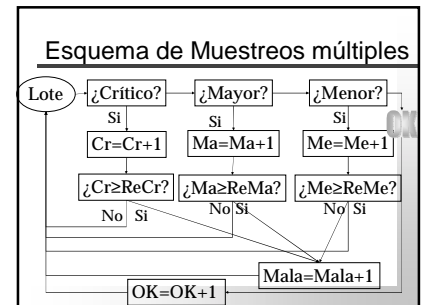
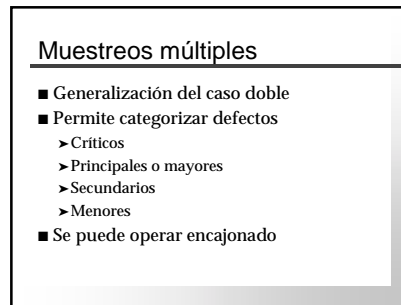
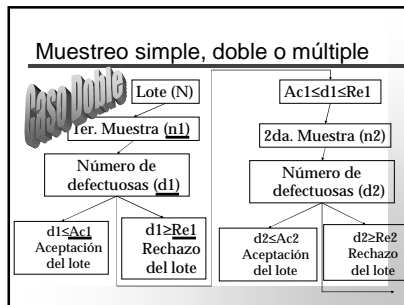
Activa 2002

Normal, Rigurosa o Reducida

- Reducida
 - Proveedores con antecedentes
 - Calidad estable y holgada
- Normal
 - Opción por defecto
- Rigurosa
 - Proveedores nuevos o con mala calidad
 - Mismo costo que Normal
 - Mayores exigencias

Muestreo simple, doble o múltiple





Recapitulando

- ✓ Control en una Cadena de Producción
 - ✓ ...
 - ✓ Implementación
 - ✓ Producción interna
 - ✓ Producción externa
 - ✓ Verificación
 - Plan de aceptación por Muestreo
 - Por Atributos
 - Por Variables
- Conceptos de ISO 9000



Muestreo para Atributos

- Notación: Atributos ~ Categórico
- Es lo más fácil
- Estándares bien establecidos
 - MIL STD 105D
 - Dodge-Romig
 - Phillips
- El cómo controlar, dónde, etc. debe estar especificado en contratos proveedor/comprador

MIL STD 105D

- Fue pionera (2da guerra)
- Se basa en el NCA
- Define y establece *niveles*
 - Generales
 - I: reducido (hay evidencia de buena calidad)
 - II: normal (para producto nuevo)
 - III: riguroso (hay evidencia de mala calidad)
 - Especiales
- Los niveles suben y bajan
- Hay criterios para cambiar de nivel1

¿Cuándo cambiar de nivel?

De	A	Criterio
I	II	Hay un rechazo
II	III	Dos de cinco lotes consecutivos rechazados
III	II	Cinco lotes consecutivos sin rechazo
II	I	Diez lotes consecutivos sin rechazo

¿Cómo se usa?

- Está organizada en tablas T1 y T2
- T1: Dado el nivel (I, II o III) y tamaño de lote N → sale un código C
- T2: Dado C → sale tamaño de muestra n
- T2: Dado C y el NCA → salen niveles de aceptación Ac y rechazo Re

Procedimiento Dodge-Romig

- Utiliza Calidad Limite CL
- Busca minimizar el # inspecciones
- Contempla muestreo simple y doble
- También con dos tablas
- T1: Dado CL, el tamaño de lote N y una estimación de \bar{x} → salen (n, Ac, LCMS)
- Alternativamente:
- T2: Dado LCMS, el tamaño de lote N y una estimación de \bar{x} → salen (n, Ac, CL)
- Problema: requiere estimar \bar{x} !

Procedimiento Philips

- Basado en PRC
- Busca igualar costos de muestreo y de no calidad
- Contempla muestreo simple y doble
- Dado el tamaño de lote N y una estimación de \bar{x} → salen (n, Ac)
- Problema: requiere estimar \bar{x} !

Muestreo para Variables

- Notación: Variables ~ Cuantitativo
- Es más difícil (más capacitación, tiempo, etc.)
- Presupone variables continuas
- Da más info → muestras más pequeñas que con atributos
- Estándares bien establecidos (MIL STD 414, etc.)
- Hay también inspección reducida, normal y rigurosa
- Desventaja: Requieren conocer la pdf
 - ¡Adivinaron! Se asume normal
- Las fases son: muestreo, medición y análisis estadístico

Por más detalles...

Nos remitimos a la literatura especializada ☺

Plan

- ✓ Control en una Cadena de Producción
- Conceptos de ISO 9000



Introducción

- Evolución del Concepto de Calidad

Requisitos del producto →

Filosofía y forma de gestión

- Implantación: tres etapas

1. Programa básico de Calidad

2. Calidad en procesos

3. Modelos de Excelencia

Objetivos:
Sensibilizar
Motivar
Avanzar a los siguientes niveles
Familia ISO 9000

Ariza 2002

¿Porqué ISO 9000?

- Estándar internacional
- Usado ampliamente
 - Bienes, servicios, etc.
- Requerimientos cruzados
- Gestión de Calidad → ISO 9001:2000
 - "Fácil" para manufactura
 - Más "difícil" en otros casos
 - Existe Manual para Cartografía en Europa (2000)
- Modelos de Excelencia → ISO 9004:2000

Familia ISO 9000

- 9000:2000 Establece los fundamentos de los SG y la terminología
- 9001:2000 Especifica los requisitos para los SGC. Aplicable a contratos. Establece requisitos mínimos
- 9004:2000 Amplía objetivos de la 9001. No es aplicable a contratos. Amplía requisitos mínimos
- 19011:2000 Da orientación relativa a auditorías de GC y medioambiental

ISO 9000 y yo

- ¿Qué puede hacer Ud.?

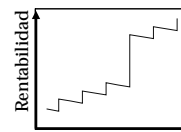
- Implementar internamente un sistema de Calidad sin mencionar al estándar
- Usar e implementar internamente el estándar
- Usar e implementar el estándar, y además certificarse ISO 9000

- Certificación

- Implica reconocimiento externo
- Implica revisión regular (cada seis meses)

- Servicio disponible en muchos países

Rentabilidad vs. mejora



Mejora
Caso de Mejora Continua
Caso de Reingeniería
y Reingeniería

Requisitos SGC ISO 9000:2000

- Sistema de Gestión de Calidad
- Responsabilidad de la Dirección
- Gestión de los Recursos
- Realización del Producto
- Medición, Análisis y Mejora

Vamos uno por vez

Sistema de Gestión de Calidad

- Requisitos generales
- Requisitos de la documentación

Responsabilidad de la Dirección

- Compromiso
- Enfoque al cliente
- Política de Calidad
- Planificación
- Responsabilidad, autoridad y comunicación
- Revisión por la Dirección



Gestión de los Recursos

- Provisión de recursos
- Recursos humanos
- Infraestructura
- Ambiente de trabajo



Realización del Producto

- Planificación de la realización del producto
- Procesos relacionados con el cliente
- Diseño y desarrollo
- Compras
- Producción y prestación del servicio
- Control de los dispositivos de seguimiento y medición



Medición, Análisis y Mejora

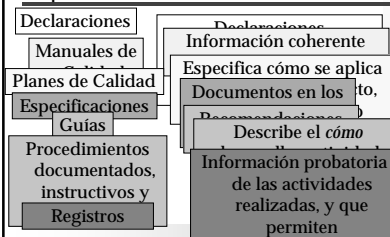
- Generalidades
- Seguimiento y Medición
- Control de producto no conforme
- Análisis de datos
- Mejora



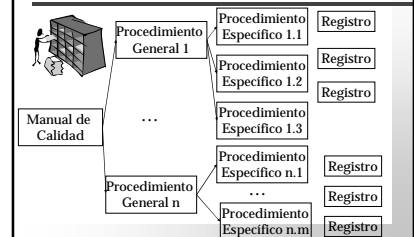
Sistema Documental

- Idea madre:
 - "...decir lo que se va a hacer..."
 - Para luego "...hacer lo que se ha dicho..."
- Incluir *todo*
- Documentación permite *comunicar* y dar coherencia a la *acción*
- Es el trabajo más arduo
- Varios tipos de documentos

Tipos de documentos



Cómo se clasifican



Documentos: Quién hace qué

Documento	Nivel de decisión	Quién participa
Manual de calidad	Estrategias de la empresa	Dirección
Procedimientos generales	Organización de procesos	Mandos Intermedios
Procedimientos específicos	Instrucciones de trabajo	Personal de base

Manual de Calidad

- Documento básico porque:
 - Será el primero solicitado
 - Es general, y menciona *todo*
 - Refleja el compromiso de la Dirección
- Documento vivo, revisado anualmente
- No debe entrar en detalles
- Estructura y contenido normalizados

Procedimientos generales

- Desarrollan lo que se perfila en el MC
 - PG de Revisión de Contrataciones
 - PG de Inspección, Ensayo y Pruebas finales
 - ...
 - PG de Formación sobre Calidad
- Debe describir acciones, decisiones, insumos (recursos), criterios, secuencias, referencias, etc.

Procedimientos específicos

- Son muy, muy propios de la empresa
- Ejemplos cartográficos:
 - PE para comprobar equipos de medición
 - PE para la realización de vuelos aerofotogramétricos
 - PE para la digitalización de documentos
 - PE para la clasificación de imágenes de satélite
- Core business

Registros

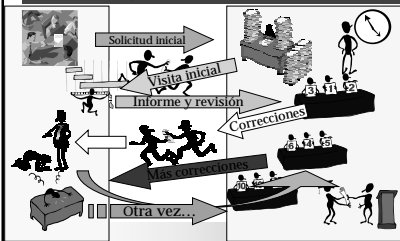
- Comprobantes o evidencias objetivas
- Validan lo hecho, o ayudan a mejorar
- Uso en auditorías internas y/o externas
- ISO 9001:2000 requiere registro de:
 - Recopilación, indexación, archivo, almacenaje, mantenimiento,...etc.
 - Idem para registros del proveedor

¿Quién controla todo esto?

- Auditorías internas
- Auditorías externas
- Certificación externa
- Características
 - Arrancan con el material escrito existente
 - Autorizadas por la Dirección
 - Auditores con amplia formación
 - Participación directa de personal auditado
 - Informe conocido y refrendado por todos
 - Conclusiones se elevan a la Dirección *siempre*

¿Quiere ISO 9000?

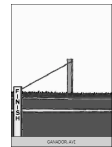
¿Cómo se procesa la Certificación?



Experiencias con ISO 9000

- IGN Francia, OS UK (~1997)
- Hay acuerdo en que ayuda, pero...
 - Aumenta la burocracia interna (papeleo)
 - Hay mayores costos de producción (nuevas tareas)
 - Resistencia al cambio
 - Más difícil subcontratar
 - ¡No mejora calidad del producto final!
- Difícil balance costo/beneficio
- Paso imprescindible
- Hay manual de CERCO (2000)

¡Esto fue todo!



Módulo 5: Manteniendo la Calidad

Carlos López Vázquez

carlos.lopez@ieee.org