International Center for Tropical Agriculture
Since 1967 Science to cultivate change

Limpieza de base de datos
errores comunes

Octubre 29, 2019

Hugo Andrés Dorado B.

Juan Camilo Rivera

h.a.dorado@cgiar.org

Contenido

- Estructura ideal de una base de datos.
- Selección de variables 1.
- Errores de formato.
- Errores en las unidades.
- Errores en los caracteres.
- Conversión de variables.
- Puntos atípicos.
- Selección de variables 2.

Estructura ideal de una base de datos (Recomendación 1)

Cada fila representa una unidad de análisis y cada columna representa una variable.

	A	B	C	D	E
1	ID	Sowing_Date	Harvest_Date	Variety	Yield
2	RC61_2008_989	2008-03-07	2008-07-05	ACARIGUA	6700
3	RC62_2010_207	2010-07-22	2010-11-25	ACD 2526	9125
4	RC62_2011_275	2011-03-11	2011-07-15	ACD 2526	6375
5	RC62_2012_361	2011-09-08	2012-01-12	ACD 2526	6875
6	RC62_2011_303	2011-04-25	2011-08-29	ACD 2528	7500
7	RC62_2011_213	2010-08-30	2011-01-03	ACD 2540	6563
8	RC62_2011_274	2011-03-09	2011-07-13	caracoli	6250
9	RC62_2010_76	2009-12-19	2010-04-24	CHICALA	5600
10	RC62_2011_336	2011-08-06	2011-12-10	CHICALA	4625
11	RC62_2011_345	2011-08-22	2011-12-26	CHICALA	4687
12	RC62_2011_348	2011-08-23	2011-12-27	CHICALA	5163
13	RC62_2012_372	2011-09-14	2012-01-18	CHICALA	6875
14	ENA_2007a_106386	2007-02-21	2007-07-01	CIMARRON BARINAS	6937.5
15	ENA_2007a_100234	2007-03-21	2007-07-25	CIMARRON BARINAS	7500
16	ENA_2007a_102633	2007-04-14	2007-09-25	CIMARRON BARINAS	8187.5
17	ENA_2007a_101504	2007-05-14	2007-10-09	CIMARRON BARINAS	8000
18	ENA_2007a_100400	2007-05-26	2007-10-06	CIMARRON BARINAS	5187.5
19	ENA_2007a_100150	2007-05-26	2007-10-13	CIMARRON BARINAS	7812.5
20	ENA_2008a_101504	2008-03-01	2008-07-02	CIMARRON BARINAS	6562.5
21	ENA_2008a_100234	2008-04-29	2008-09-09	CIMARRON BARINAS	7000

Asegurarse de crear o identificar un ID que le permita conectar las bases de datos

Estructura ideal de una base de datos (Recomendación 2)

Crear un diccionario de variables

Practica

	Nombre corto	**Dato de la practica**	Tipo	Opciones pensados	
Preparacion de la parcela	**fechaTrabajo**	**Fecha de trabajo**	Fecha		
	tipoPreparacion	tipo de preparacion		Labor + número de pases: Subsolador, cincel, arado, rastra, rastrillo, micronivelación, embalconado o encamado.	
	profTrabajo	Profundidad de trabajo	Numero	[30 - 100](cm)	
	manejoRastrojos	Manejo de rastrojos		ninguno, quema, integracion al suelo, picados (desbrozadora o combinada)	
Siembra	**fechaSiembra**	**Fecha de siembra**	Fecha		
	tipoSiembra	Tipo de siembra(maqinaria)		Convencional, directa, manual.	
	semillas	Semillas / ha	Número	Número	
	tipoMaterial	Tipo de material		Variedad, Hibrido, OGM, semilla campesina	
	colEndospermo	Color del endospermo		Blanco o amarillo	
	materialGenetico	Material genetico (nombre)		Lista de los materiales usados en Colombia (los mas sembrados y otros)	
	semillaTratada	Semillas tratadas ?		SI/NO	
	producto	Con que producto		Fungicidas, insecticidas, otro	
Datos generales	objetRendimento	Objetivo de rendimiento	Numero	(kg/ha)cuánto espero del cultivo ?	
	cultivAnterior	Cultivo anterior		Lista de cultivos de Colombia	Soya , arroz, algodón , maíz, sorgo, pastos , otros...
	drenajeParcela	Se hace drenaje en la parcela		SI/NO	

Como mínimo se sugiere, por cada variable:

- Nombre corto
- Nombre completo
- Unidad de medida
- Rango [Max - Min], o posibles categoríaas

Estructura ideal de una base de datos (Recomendación extra)

	Drenaje	variedad	tratSemilla	tipLab	ano	insecticida	fungicida	Limoprc	Arenaprc	DRAvg	RSAvg	RHsd	rendimiento
80246_1	D4	OTRAS	S1	LabCeroM	2016	2	0	20	52	11.53848	449.2265	3.655242	7
94278_1	D2	OTRAS	S0	AgrCons	2016	1	0	20	36	11.87806	447.6698	3.832183	6.2
53520_1	D4	OTRAS	S0	LabConvTr	2015	1	0	20	52	11.36212	426.8253	4.657267	2
74603_1	D2	OTRAS	S0	LabConvTr	2016	1	1	20	36	10.42686	475.9483	4.986175	3.2
74604_1	D2	OTRAS	S0	LabCeroM	2016	1	1	20	36	10.42686	475.9483	4.986175	3.5
81785_2	D4	CLTHW11(	S1	AgrCons	2016	0	0	26	48	12.1955	438.4033	4.41719	0.02
74327_1	D4	OTRAS	S1	LabConvTr	2016	2	0	27	56	11.51007	460.2183	3.525872	5.32
74328_1	D4	OTRAS	S1	AgrCons	2016	2	0	27	56	11.49896	459.8823	3.498673	6.09
74234_1	D4	CLTHW11(	S1	LabConvTr	2016	1	0	26	48	11.50372	459.3266	3.484351	5.81
74235_1	D4	CLTHW11(	S1	AgrCons	2016	1	0	26	48	11.50372	459.3266	3.484351	5.65
74340_1	D4	CLTHW11(	S1	LabConvTr	2016	1	0	26	48	11.51433	460.6882	3.56348	5.11
74341_1	D4	CLTHW11(	S1	AgrCons	2016	1	0	26	48	11.51433	460.6882	3.56348	4.98
74344_1	D2	CLTHW11(	S1	LabConvTr	2016	1	0	20	36	11.4932	460.5205	3.602382	5.02
74345_1	D2	CLTHW11(	S1	AgrCons	2016	1	0	20	36	11.4932	460.5205	3.602382	5.02
87662_1	D4	OTRAS	S0	AgrCons	2016	0	0	27	56	11.49163	434.1986	3.595167	4
81785_1	D4	CLTHW11(	S1	AgrCons	2016	0	0	26	48	12.1955	438.4033	4.41719	3.5
90733_1	D4	CLTHW11(	S0	AgrCons	2016	1	0	26	48	11.52554	440.353	3.653417	5
15081_1	D4	OTRAS	S0	LabConvTr	2012	0	0	22	56	13.49568	353.7543	7.181806	3.7
7959_1	D4	CRIOLLO	S0	LabConvTr	2012	1	0	26	44	12.33696	377.7646	9.6143	2.8
7960_1	D4	CRIOLLO	S0	AgrCons	2012	1	0	26	44	12.17765	377.0034	9.476959	4
8532_1	D4	CRIOLLO	S0	LabConvTr	2012	1	0	26	44	12.5716	383.8432	9.741766	2

Cualitativas (X)

Cuantitativas (X)

Resupuesta (Y)

Selección de variables 1

Remover variables que no tiene ningún sentido mantener tales como: datos personales,

ID	FECHA_ENCUESTA	NOM_PROD	CEDULA	LUGAR_EXPE	TEL_MOVIL	TEL_FIJO	CORREO ELECTRONICO	NOM_FINCA	DIR_RESIDENCI
212		ANDRES EDUARDO MEJIA HERNANDEZ	15388154	NA	NA	54101	amejiah2@gmail.com	EL GUARANGO	NA
213		BAIRO DE JESUS CIRO RESTREPO	15375159	NA	NA	55302	NA	BAIRO CIRO	NA
214		BERNARDO DE JESUS HENAO ESCOBAR	71141593	NA	NA	56403	NA	EL MIRADOR	NA
215		BERTHA INES CASTAÑO MUÑOZ	21952916	NA	NA	54218	NA	VILLA ELISA	NA
216		BERTULFO ROMAN FLOREZ	3558284	NA	NA	56412	NA	PARAJE PANTALIO	NA
217		C.I. FRUTY GREEN S.A.	900155227	NA	NA	54124	gio@une.net.co	EL CEBADERO	NA
218		CARLOS ALBERTO BEDOYA BEDOYA	15384516	NA	NA	NA	NA	LA FURA1	NA
219		CARLOS ANDRES ROMAN	15385357	NA	NA	NA	NA	EL MORRITO	NA
220		CARLOS MARIO SALAZAR BERMUDEZ	71724382	NA	NA	NA	monicamg83@hotmail.com	NA	NA
221		CONSTRUCCIONES Y FINCAS (RICARDO ECH	800041356-4	NA	NA	26650	ricechavarria@hotmail.com	LA PRADERA	NA
222		DIANA CAROLINA RAMIREZ GALVIS	1040031578	NA	NA	56409	caroramirez12@hotmail.com	LA SAMARIA	NA
223		DIEGO FERNANDO SALAS 3393200 E1 132	71317277	NA	NA	41635	diegosalsas@hotmail.com	LOS FAROLES	NA
224		DIEGO LEON TOBON BEDOYA	15389108	NA	NA	56253	NA	EL SALADERO 2	NA
225		EDGAR DARIO TOBON TOBON	15382473	NA	NA	56253	NA	EL SALADERO	NA
226		EDUARDO ANTONIO ECHEVERRI ANGEL	8231045	NA	NA	54100	malejaem@yahoo.es	PLAYITA LINDA	NA
227		EFRAIN DE JESUS BEDOYA BEDOYA	71555462	NA	NA	54112	NA	LA FORTUNA	NA
228		EFRAIN DE JESUS HENAO ESCOBAR	71555627	NA	NA	56403	NA	LA GEMELA	NA
229		EFRAIN DE JESUS VILLEGAS BEDOYA	3558570	NA	NA	54107	NA	NA	NA
230		ELENA GONZALEZ AYALA	27450716	NA	NA	54103	NA	NA	NA
231		ELKIN DE JESUS OSPINA MONTOYA	3327938	NA	NA	26866	NA	SAN MIGUEL	NA
232		ENRIQUE MONTOYA PELAEZ	6789631	NA	NA	32658	NA	NA	NA
233		ERIKA MARIA CASTAÑO GOMEZ	43878498	NA	NA	56413	NA	LEJANIAS	NA
234		FRANCISCO ALBERTO VARGAS MORALES	15349653	NA	NA	54125	fvargas74@yahoo.com	LA MERCED	NA
235		FABIAN CADAVID ORTIZ	NA	NA	NA	54140	fabianacadavid@hotmail.com	LAS COLINAS	NA
237		GABRIEL EDUARDO BOTERO MUÑOZ	8308879	NA	NA	54105	NA	LA LUCIA	NA
238		GLADIS DE JESUS ZAPATA DE CASTAÑO	32340043	NA	NA	23573	facano@gmail.com	PROVIDENCIA	NA
239		GLORIA ISAZA DE RESTREPO	21272220	NA	NA	41361	mdomaglo@une.net.co	EL MANANTIAL	NA
241		GUSTAVO ALONSO BEDOYA BOTERO	71555442	NA	NA	56409	NA	NA	NA
242		GUSTAVO RESTREPO ACEVEDO	3341118	NA	NA	54206	NA	ALTOS DE MAZARELO	NA
243		HECTOR DANIEL HENAO BOTERO	3558496	NA	NA	5411014	NA	LOS SOLECITOS	NA

Errores de formato

Fechas

	O	P	Q	R	S	T
		ManeraSiemb	FechaGerminacion	FechaCosecha	CrecimientoTotal	Area
711082		Chorrillo	6/5/2016	10/5/2016	122	
646509		Chorrillo	7/10/2016	11/9/2016	122	
500618		Espeque	8/10/2018	12/6/2019	483	
085106		Espeque	8/29/2018	12/19/2018	112	
496744		Espeque	9/14/2018	1/8/2019	116	
021449		Espeque	8/15/2018	12/12/2018	119	
534115		Espeque	8/17/2018	12/15/2018	120	
756709		Espeque	8/31/2018	9/27/2018	27	
235164		Chorrillo	7/6/2016	11/5/2016	122	
569002		Chorrillo	6/27/2016	10/22/2016	117	
905146		Chorrillo	6/27/2016	10/23/2016	118	
401172		Chorrillo	6/24/2016	10/17/2016	115	
086117		Chorrillo	6/5/2016	10/5/2016	122	
436118		Chorrillo	6/7/2016	10/1/2016	116	
038814		Espeque	6/10/2016	10/12/2016	124	
544835		Espeque	6/17/2016	10/20/2016	125	

Fechas

FECHA DE Germinacion	FECHA DE COSECHA	Rendimiento unitario qq/mz
6/2/2016	5-Oct-16	62.7
6/2/2016	25-Oct-16	39.9
6/10/2016	8-Oct-16	128
6/14/2016	10-Oct-16	80.49
6/6/2018	6-Oct-16	132.99
6/10/2016	8-Oct-16	74.91
6/8/2016		99
Ago.04/2016	8-Dec-16	123.75
2016		116.05
2016		118.184
Jul.16/2016	Nov.23/2016	137.5
Jul.06/2016	Nov.15/2016	53.999
Jun.27/2018	Nov.15/2016	41.316
Jun.27/2016	Nov.16/2016	109.197
Jun.24/2017	Nov.10/2016	34.9965
Jul.07/2016	Dic.01/2016	150.0015
Jun.10/2016	Oct.12/2016	139.348
Jun.17/2016	Oct.20/2016	158.4

Coordenadas mal registradas

37	5°08'27.5"	-75°54'31.3"	RISARALDA	APIA
38	5°08'42.3"	-75°55'02.2"	RISARALDA	APIA
39	5°78'41.0"	-75°05'02.8"	RISARALDA	APIA
40	5°67'16.8"	-75°04'17.0"	RISARALDA	APIA
41	5°08'17.8"	-75°54'18.4"	RISARALDA	APIA
42	5°09'41.8"	-75°55'26.4"	RISARALDA	APIA
43	5°09'41.6"	-75°55'26.1"	RISARALDA	APIA
44	5°09'35.3"	-75°55'10.0"	RISARALDA	APIA

Errores en las unidades

Unidad de producción

ID_FIN	PRODUCTOR	RESIEMBRA_PLATANO	PN_ANO_PLATANO	UNIDAD_PN_PLATANO
6	ELIODORA DONAD CHAUTACA	10	3000	KG
7	ERMYS ROQUEME ALMANZA		7000	KG
13	DARLIS FUENTES PEÑATE		7200	KG
14	CRISTINA HERNANDEZ		500	KG
16	GERARDO ENRIQUE MUÑOZ VERA	20	30000	KG
17	ANCIZAR MUÑOZ VERA	20	52000	KG
19	NELSON FABRA DÍAZ		8500	KG
20	MARIA NICOMEDES DIAZ	30	5000	KG
22	FREDY SALEME GONZALEZ		3000	KG
45	JULIO CANSADO M.		520	KG
86	OMAR JIMENEZ OLAVARRIA		1000	KG
88	ANA ROSA HERRERA GONZALEZ		4800	KG
89	NICOLA REDONDO PAHECO		35	KG
90	VICENTE NOVIL BOLAÑO		140	KG
91	NICOLAS BRITO IGUARAN		9600	KG
96	DANIEL AMAYA		100	KG
97	ALBERT REDONDO	20	3000	KG
99	GIOVANNI MILTON DEL PRADO ANAYA	50	3000	KG

Unidad de área

25	26	27	28
NO_LOTES_MANGO	NO_ARBO	AREA_MA	UNIDAD_MEDID
4	25	0.001	HECTAREAS
3.5	2	0.25	HECTAREAS
1	5	0.5	HECTAREAS
1	50	0.5	HECTAREAS
1	25	0.5	HECTAREAS
2	400	4	HECTAREAS
1	50	1.5	HECTAREAS
1	35	0.6	HECTAREAS
1	105	0.5	HECTAREAS
1	200	2	HECTAREAS
1	40	0.5	HECTAREAS
1	1300	1	HECTAREAS
1	2	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
NA	NA	NA	DISPERSOS
2	800	5.76	HECTAREAS
220	220	2	HECTAREAS
NA	700	NA	HECTAREAS
NA	NA	8	FANEGADAS
2	160	1.5	HECTAREAS
NA	NA	NA	HECTAREAS
3	450	5	HECTAREAS
5	NA	6	HECTAREAS
2	600	7	HECTAREAS
NA	NA	NA	HECTAREAS

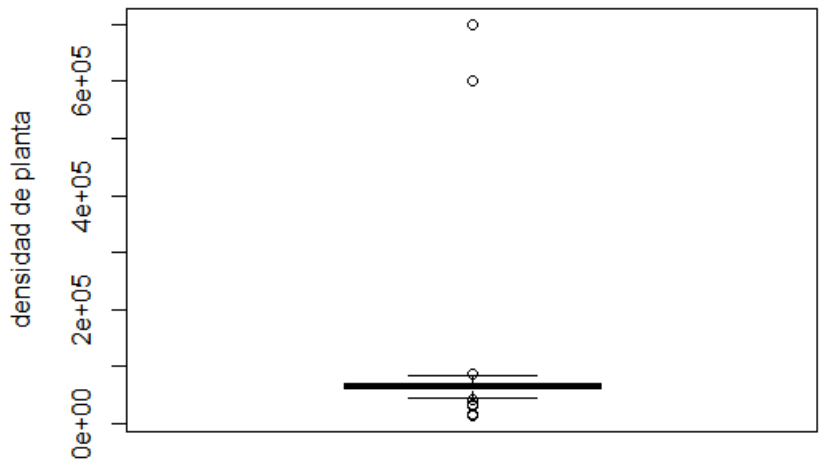
Errores en los caracteres

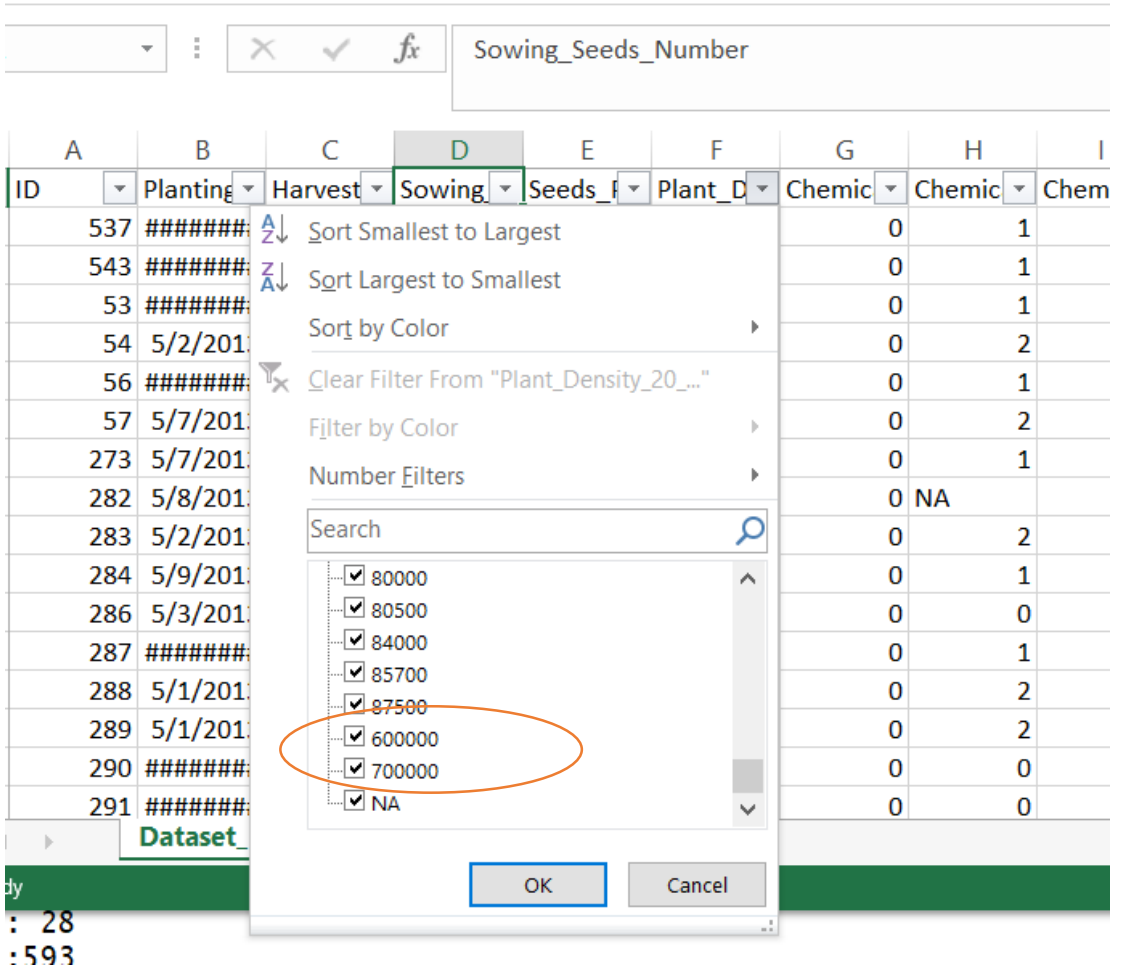
VARIEDAD	Manera DE SIEMBR	FECHA DE Germinaci
INTA Fortaleza Secano	Chorrillo	junio
INTA L9	Chorrillo	
INTA Fortaleza Secano	Chorrillo	
INTA L9	Chorrillo	junio
Inta L9	Maquinaria	7/22/2016
Inta L9	Maquinaria	7/18/2017
INTA Dorado	Bueyes	Julio
Inta F. Secano	Bueyes	8/17/2016
INTA L9	Espeque	INTA L9
INTA Fortaleza Secano	Espeque	INTA Fortaleza Secano
Inta San Juan	Se Perdio	
INTAL8,	Espeque	10/10/2016
Inta L9	Espeque	10/12/2016
INTA Chinandega	Chorrillo	10/11/2016
INTA Chinandega	Espeque	10/8/2016
INTA L8, INTA L9 INTA Fortaleza Secano	Chorrillo	No Siguieron para 2017
INTA L9	Chorrillo	6/10/2017
Inta L9	Chorrillo	6/15/2017

Conversión de variables.

- Fechas a días julianos.
- Categorías poco repetidas colocarlas en una categoría como otras.
- Variables como fertilización, número de monitoreos, variables climáticas, deben ser resumidas.

Puntos atípicos





Selección de variables 2.

- Remover variables irrelevantes.
- Remover variables redundantes.

Thank you!

CIAT 50
1967-2017

WE'RE PROUD TO
HAVE CELEBRATED 50 YEARS
OF AGRICULTURAL RESEARCH
FOR DEVELOPMENT

International Center for Tropical Agriculture - CIAT

Headquarters and Regional Office
for South America and the Caribbean

+57 2 445 0000
Km 17 Recta Cali-Palmira
A.A. 6713, Cali, Colombia

ciat@cgiar.org
ciat.cgiar.org

CGIAR
CIAT is a CGIAR Research Center