

Session 3 Synthèses de données initiales

Il est capital d'utiliser une procédure systématique pour nettoyer les données. Cela permet de suivre précisément les éventuelles modifications apportées et permet aux autres de comprendre comment vous êtes passé des données brutes aux données nettoyées que l'on utilisera pour l'analyse.

Pour faire suite à la section précédente, maintenant que vous avez enregistré une copie des données d'origine et vérifié les ID, nous allons produire des synthèses des données brutes. Cela vous permettra de comprendre les données et d'organiser et stocker les éventuelles modifications apportées.

Dans cette optique, une série d'étapes va vous être présentée. Vous pouvez suivre ces étapes en utilisant les données d'entraînement fournies.

NB Nous utiliserons le 'classeur de nettoyage de données' et 'données 2' pour débiter cette session.

3.1 Procédure de production des synthèses de données

Le Classeur de nettoyage des données contient des modèles de production des synthèses de données dans les feuilles intitulées « Synthèse école brute » et « Synthèse élève brute ».

NB Vous pouvez copier et renommer les feuilles si vous disposez d'un plus grand nombre d'ensembles de données.

Ici nous allons montrer comment compléter la feuille de synthèse à l'aide des données d'école.

Le processus :

1. Commencer par saisir les informations basiques de l'ensemble de données en haut à gauche de la feuille de calcul.

	nom de l'ensemble de données	
2		
3	Nombre de lignes	
4	Nombre de colonnes	
5	Identifiant unique	

2. Pour compter rapidement le nombre de lignes, mettre en surbrillance la première colonne de données et noter le nombre qui s'affiche en bas à droite de la page :

Ligne école	Date de l'enquête	Région	District	Code district	Latitude arrivée	Longitude arrivée	Latitude départ	Longitude départ
1	03/07/2015	AMHARA	BELB WUHA	120	12.18.550	037.45.04	12.118.542	037.45.075
2	03/09/2015	AMHARA	DAWA CHEFE	117	10.46.084	039.51.800	10.46.084	039.05.799
3	03/03/2015	AMHARA	KEMISE	117	10.52.371	039.51.607	10.52.370	039.05.670
4	03/11/2015	SNNP	GURAFERDA	123	06.83.725	035.29.902	06.83.725	35.29.872
5	03/12/2015	SNNP	GURAFERDA	123	06.86.817	035.34.801	06.87.925	035.25.734
6	03/03/2015	SNNP	MIZAN TOWN	122	06.99.501	035.58.899	06.99.506	035.58.890

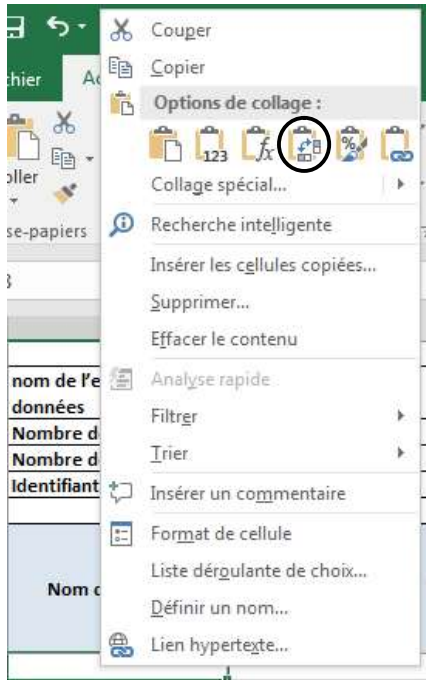
3. Vous pouvez compter le nombre de variables de façon similaire, mettre en surbrillance la première colonne de données et noter le nombre qui s'affiche en bas à droite de la page :

Longitude départ	Nom école	Code école	Nom du directeur	Nombre de garçons	Nombre de filles
037.45.075	BELB W	101	WASIHUN WENDIMAGEGNE	129	104
039.05.799	BILLECHA	102	TARIKU	1002	137
039.05.670	SERTIE	103	REGASSA WAKIYA	78	127
35.29.872	OTWEA No 2	104	Mengistu Erose	112	
035.25.734	ORENITA	105	AYENEV BERELE	146	108
035.58.890	MESEGAN AKADAME	106	TATEK KEBEDE	99	116

4. Saisir le nom de la variable de la colonne d'identifiant unique dans la case d'identifiant unique, dans ce cas il s'agit de « Code école (XXX) ».
5. **Noms de variable** – Copier les en-têtes de variables de l'ensemble de données vers la feuille de synthèse brute :
Mettre en surbrillance la ligne qui contient les noms de variable et copier (ctrl+c ou clic droit puis copier).

Date de l'enquête	Région	District	Code école	Longitude arrivée	Latitude départ	Longitude départ	Nom de l'école	Code de l'école	Nom du directeur	Nombre de filles	Nombre de garçons	Nombre total des élèves	Les élèves de votre établissement ont-ils reçu un traitement vermifuge l'année dernière (0 = N, 1 = O)
03/07/2015	AMHARA	BELB WUHA	120	12.18.550	037.45.04	12.118.542	BELB W	101	WASIHUN WENDIMAGEGNE	129	104	233	0
03/09/2015	AMHARA	DAWA CHEFE	117	10.46.084	039.51.800	10.46.084	BILLECHA	102	TARIKU	1002	137	239	1
03/03/2015	AMHARA	KEMISE	117	10.52.371	039.51.607	10.52.370	SERTIE	103	REGASSA WAKIYA	78	127	205	0
03/11/2015	SNNP	GURAFERDA	123	06.83.725	035.29.902	06.83.725	OTWEA No 2	104	Mengistu Erose	112		112	1
03/12/2015	SNNP	GURAFERDA	123	06.86.817	035.34.801	06.87.925	ORENITA	105	AYENEV BERELE	146	108	254	1
03/03/2015	SNNP	MIZAN TOWN	122	06.99.501	035.58.899	06.99.506	MESEGAN AKADAME	106	TATEK KEBEDE	99	116	215	1

Faire un clic droit sur la cellule ci-dessous « Nom de variable » dans la feuille de synthèse (A8) et dans « Options de collage » sélectionner l'option Transposer.



6. **Facteur ou nombre ?** – Cette colonne doit être remplie à la main. En général, une variable est un facteur si elle possède un choix limité d'options pour classer quelque chose (par ex., oui/non ou fille/garçon) et un nombre sert à quantifier des données types (par ex., œufs de *S. mansoni*).

7.

Facteur : nombre de niveaux Nombre : minimum	Facteur : laisser vierge. Nombre : moyenne	Facteur : laisser vierge. Nombre : maximum
---	---	---

Prendre chaque variable de votre ensemble de données et produire des synthèses des informations qu'elles contiennent.

Si la variable est un facteur, remplir uniquement la première de ces trois colonnes : « Facteur : nombre de niveaux ». S'il s'agit d'un nombre, remplir les trois colonnes, minimum, moyenne et maximum.

Niveaux

Pour trouver le nombre de niveaux, utiliser un tableau croisé dynamique :
Mettre en surbrillance vos données dans chaque colonne, sélectionner

l'onglet Insertion dans la barre d'outils en haut et cliquer sur « Tableau croisé

The screenshot shows the Excel interface with the 'Tableau croisé dynamique' task pane open on the left. The background data table is as follows:

Date de l'enquête	Région	District	Code du district (XXX)	Latitude arrivée (XXX.XXXXX)	Longitude arrivée (XXX.XXXXX)	Latitude départ (XXX.XXXXX)	Longitude départ (XXX.XXXXX)	Nom de l'école	Code de l'école (XXX)	Nom du directeur	Nombre des filles	Nombre des garçons	Nombre total des élèves	Les élèves de votre établissement ont-ils reçu un traitement vermifuge l'année dernière (O = Non, 1 = Oui)
03/07/2015	AMHARA	REB WUPHA	120	12.18.550	037.45.04	12.118.542	037.45.075	REB W.	101	WASHUN WENDIMAGEGNE	129	104	233	0
03/09/2015	AMHARA	DAWA CHEFE	117	10.46.084	039.51.800	10.46.084	039.05.799	BILLECHA	102	TARIKU	1002	197	299	1
03/03/2015	AMHARA	KEMISE	117	10.52.371	039.51.607	10.52.370	039.05.670	ERTIE	103	REGASIA WAKYA	78	127	205	0
03/11/2015	SNMP	GURAFERDA	123	06.85.725	035.29.902	06.85.725	35.29.872	OTENA No 2	104	Mengistu Dose	112	112	224	1
03/12/2015	SNMP	GURAFERDA	123	06.86.817	035.34.801	06.87.925	035.25.734	ORENTA	104	AYENEW BEBELE	146	108	254	1
03/03/2015	SNMP	MIZAN TOWN	122	06.99.501	035.58.899	06.99.506	035.58.890	MESEGAN AKADAME	106	TATEX KEBEDE	99	116	215	1

dynamique » puis « OK » (normalement c'est la première option) :

Vous obtiendrez une nouvelle feuille de calcul qui ressemble à ceci :

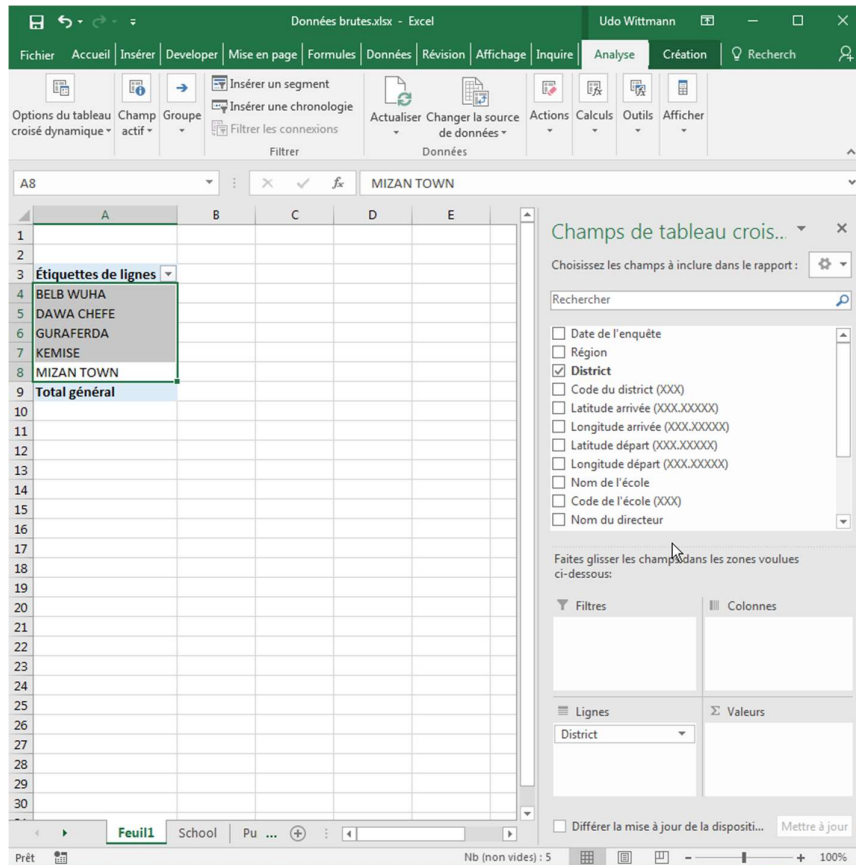
The screenshot shows the Excel interface with the 'Champs de tableau croisé dynamique' task pane open on the right. The task pane contains a list of fields to be included in the report:

- ☐ Date de l'enquête
- ☐ Région
- ☐ District
- ☐ Code du district (XXX)
- ☐ Latitude arrivée (XXX.XXXXX)
- ☐ Longitude arrivée (XXX.XXXXX)
- ☐ Latitude départ (XXX.XXXXX)
- ☐ Longitude départ (XXX.XXXXX)
- ☐ Nom de l'école
- ☐ Code de l'école (XXX)
- ☐ Nom du directeur

Below the list, there are four zones for placing the fields: Filtres, Colonnes, Lignes, and Valeurs. At the bottom, there is a checkbox for 'Différer la mise à jour de la disposition...' and a 'Mettre à jour' button.

The background shows a new worksheet titled 'Feuil1' with a grid of data. A text box in the grid reads: 'Pour générer un rapport, choisissez des champs dans la liste des champs de tableau croisé dynamique'.

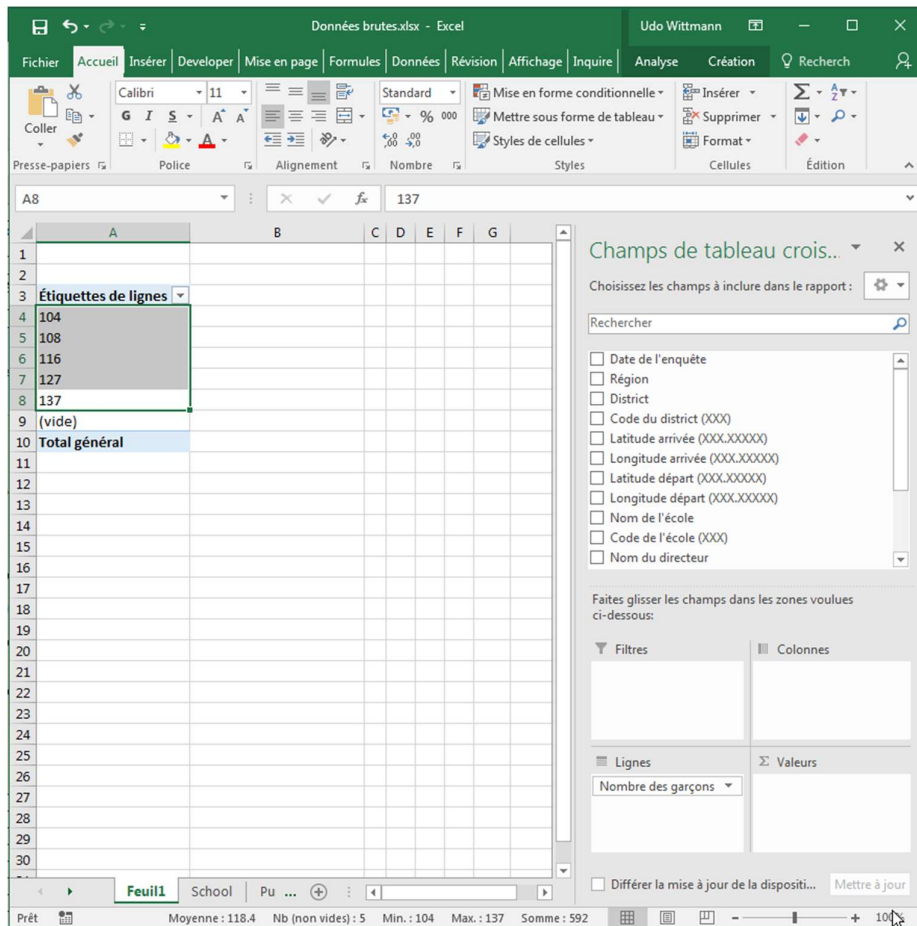
Pour voir combien de valeurs uniques existent pour chaque variable, sélectionner la variable qui vous intéresse et la faire glisser dans la fenêtre « Étiquettes de ligne ». Les valeurs uniques seront listées dans la zone de la feuille étiquetée « Tableau croisé dynamique 1 ». Vous pouvez ensuite mettre en surbrillance les valeurs uniques et noter le nombre qui s'affiche en bas à droite de la page :



Remplir la colonne appropriée de la feuille de synthèse des données. Le nombre de valeurs uniques est-il celui prévu ?

Comptes

Évaluer la variable numérique de la même manière. Sélectionner la variable et la faire glisser dans la fenêtre « Étiquettes de ligne ». Mettre en surbrillance les valeurs indiquées dans la zone Tableau croisé dynamique. Les informations en bas à droite de la page doivent afficher Moyenne, Compte (le nombre de valeurs est-il celui prévu ?), Min et Max :



Vous pouvez utiliser les valeurs en surbrillance pour remplir la feuille de synthèse des données. Utiliser la valeur Compte pour vérifier s'il manque des valeurs.

[Avant l'étape suivante, nous prendrons les données d'école et compléterons le tableau de synthèse.]

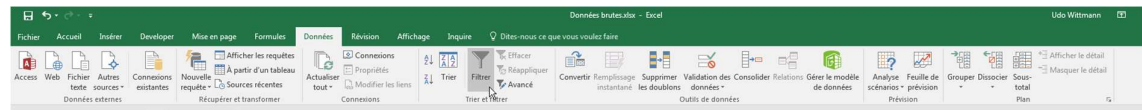
8. Nombre de valeurs manquantes

Le nombre de lignes remplies à l'étape 1 vous permet de savoir combien de valeurs il doit y avoir pour chaque variable. Utiliser le Compte indiqué précédemment pour déterminer si des valeurs manquent et saisir ce nombre dans la colonne concernée.

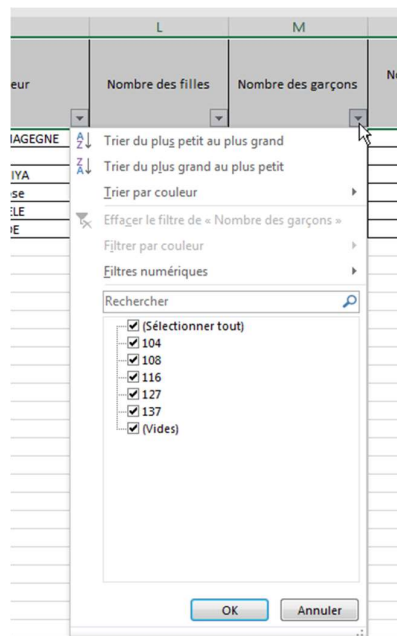
Méthode alternative vérifier les valeurs manquantes

Vous pouvez également utiliser l'outil « Filtre » pour voir s'il manque des valeurs.

Cliquer sur le numéro de ligne de la ligne qui contient les noms de variable (dans ce cas, la ligne 1). Sélectionner le menu Données dans le ruban en haut puis « Filtrer » :



Date de l'enquête	Région	District	Code du district (XXX)	Latitude arrivée (XXX.XXXXX)	Longitude arrivée (XXX.XXXXX)	Latitude départ (XXX.XXXXX)	Longitude départ (XXX.XXXXX)	Nom de l'école	Code de l'école (XXX)	Nom du directeur	Nombre des filles	Nombre des garçons	Nombre total des élèves	Les élèves de votre établissement ont-ils reçu un traitement vermifuge l'année dernière (0 = Non, 1 = Oui)
09/07/2015	AMHARA	BELB WUHA	120	12.18.550	037.45.04	12.118.542	037.45.075	BELB W	101	WASHUN WENDIMAGEGNE	129	104	233	0
09/09/2015	AMHARA	DAWA CHEFE	117	10.46.084	039.51.800	10.46.084	039.05.799	BILLECHA	102	TARIKU	1002	137	239	1
09/09/2015	AMHARA	KEMBE	117	10.52.371	039.51.807	10.52.370	039.05.670	SEITE	103	REGASSA WAKITTA	78	127	205	0
09/11/2015	SNMP	GURAFERDA	123	06.83.725	035.29.902	06.83.725	35.29.872	OTEWIA No 2	104	Mengistu Eroso	112		112	1
09/12/2015	SNMP	GURAFERDA	123	06.86.817	035.34.801	06.87.925	035.25.734	ORENITA	104	AYENEW BERELE	146	108	254	1
09/09/2015	SNMP	MIDAN TOWIN	122	06.99.501	035.58.899	06.99.506	035.58.890	MESSEGANA AKADAMIE	106	TATEK KEBEDE	99	116	215	1



9. Est-ce ce que l'on attend ?

Vérifier la synthèse de chaque variable et déterminer si elle est raisonnable pour les données recueillies (par ex., le nombre de districts est-il correct, la plage du nombre d'enfants est-elle raisonnable ?). Cette étape vous donnera une idée des points sur lesquels vous concentrer.

10. Maintenant que vous avez synthétisé l'état actuel des données, vous pouvez prendre chaque variable une à une et décider si des modifications doivent y être apportées. Veiller à noter chaque modification faite.

Si vous devez faire de nombreuses modifications, utiliser la feuille « Enquête école » ou « Enquête élève » pour les stocker. Il est recommandé d'inclure la référence de ligne ou l'ID unique de chaque variable modifiée de sorte que l'on puisse facilement retracer vos actions.

Maintenant, à vous : préparer une synthèse des données d'élève brutes

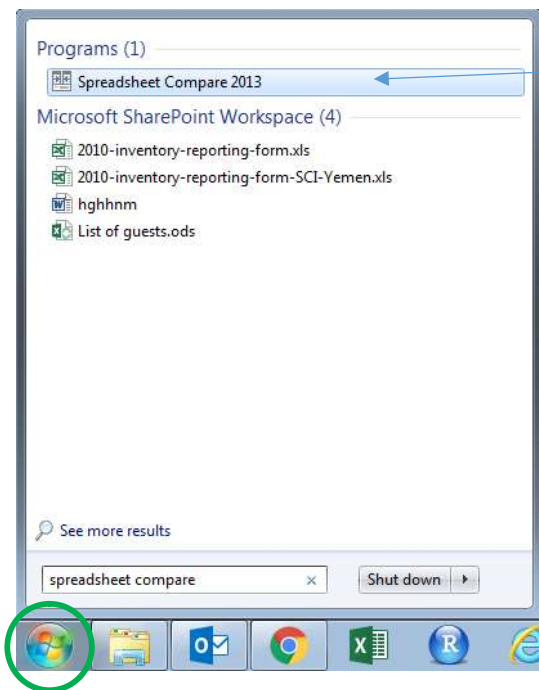
3.2 Préparer une synthèse des données d'élève.**3.3 Enregistrer vos données**

Après avoir examiné vos données et apporté les éventuelles modifications, enregistrer sous « données 3 ».

Enregistrer votre « classeur de nettoyage de données ».

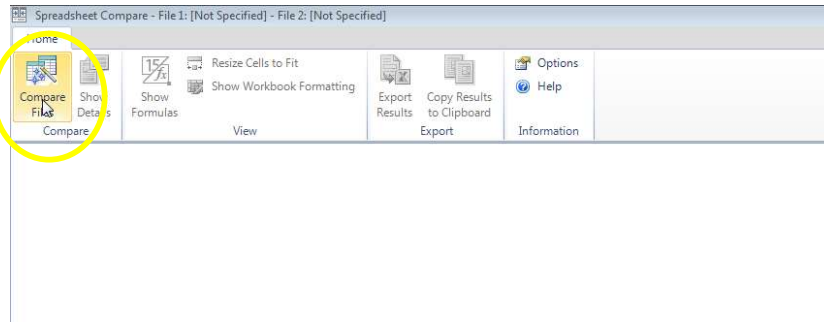
3.4 Comparer des feuilles de calcul

Pour voir facilement les modifications apportées, vous pouvez utiliser le programme « Comparer des feuilles de calcul » de MS Office.

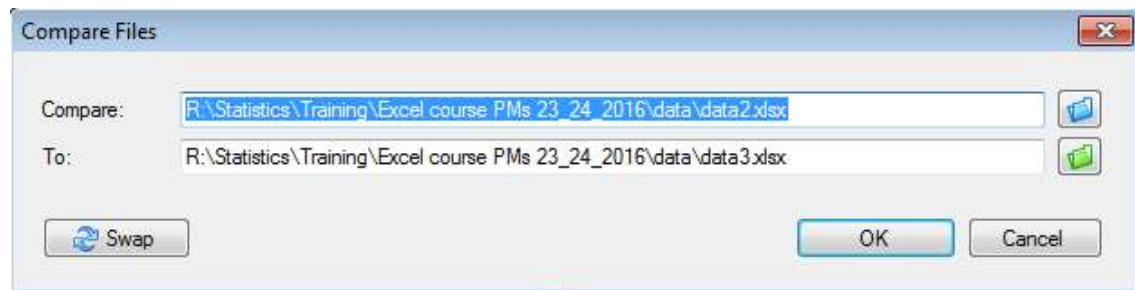


Cliquer sur le bouton Windows (entouré) et taper « spreadsheet compare » dans la case Rechercher.

Cliquer sur le bouton « Comparer des fichiers » à gauche dans le ruban menu :



Sélectionner les deux fichiers que vous souhaitez comparer dans la boîte de dialogue et cliquer sur OK.



Spreadsheet Compare affiche les deux feuilles de calcul côte à côte et met en surbrillance les cellules qui diffèrent.

School row	Survey Date	Region	District	District Cod	Arrival Lat	Arrival Lon	Departure	School No	School Cod	Headmaster	Number of	Number of gr
1	7/3/2015	AMHARA	BELB WUH	120	12.18.559	127.45.04	12.18.542	127.45.075	BELB W	101	WASHUN	120
2	9/3/2015	AMHARA	DAWAK CHE	117	13.46.084	135.51.800	13.46.084	135.51.799	BELLECHA	102	TASRU	117
3	3/3/2015	AMHARA	KEMISE	117	10.52.371	109.51.607	10.52.370	109.51.670	SERTIE	103	REGASSA	78
4	11/3/2015	SNRP	GURAFERO	123	06.83.725	135.29.802	06.83.725	135.29.872	OTENA No	104	Mengabo E	112
5	12/3/2015	SNRP	GURAFERO	123	06.88.817	135.29.801	06.87.825	135.29.794	ORIENTA	105	AYENDE B	146
6	3/3/2015	SNRP	HIZAN TO	122	06.99.551	135.58.899	06.99.506	135.58.890	MESEKANA	106	TATEK KEB	99