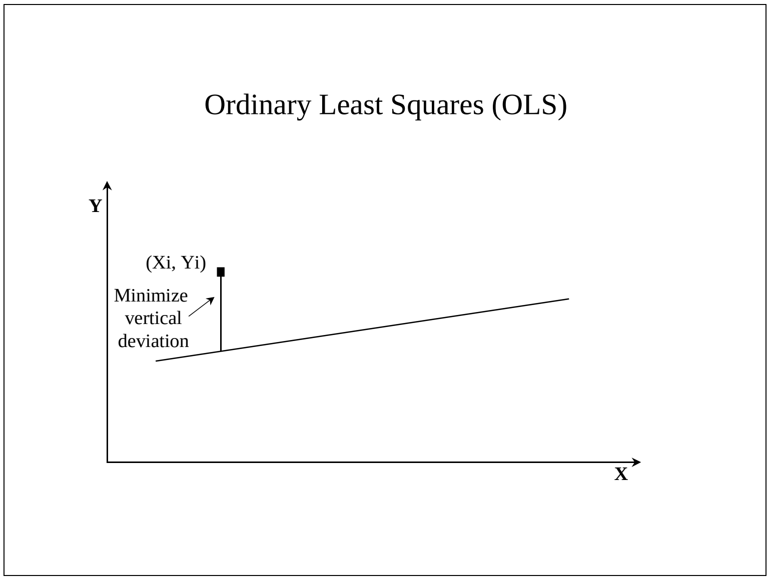
Command Reference: FillRegression()

Fill missing time series data using ordinary least squares regression

Version 10.03.00, 2012-01-08

The `FillRegression()` command fills missing data in a time series using ordinary least squares (OLS) regression and provides a variety of options for transforming the data and controlling the analysis. In OLS regression, the vertical distance from the data point to the regression line is minimized. OLS regression provides the minimum-variance estimate for a single value or observation. However, if an ensemble of points is estimated from OLS regression, the estimated values will have lesser variability than the true values.



See also the `FillMOVE2()`command, which utilizes additional variance from independent time series to determine the regression relationship, and the `FillMixedStation()` command, which automates the analysis of many time series to determine a “best estimate” filling approach. Regression can be applied only to regular interval time series. The dependent time series will be filled using the independent time series. The periods of record and output period for the time series should be verified to make sure that the time series periods overlap sufficiently. Regression relationships are developed using the analysis period for the time series and are applied to the fill period. Refer to the output statistics table, log file, and time series properties for analysis details. Several parameters are available to ensure that filling uses reasonable relationships. This command has functionality that may not be needed for simple analysis but which is useful for software testing and comparison with the `FillMixedStation()` command.

Important: TSTool does allow filled values to be flagged. However, other commands do not exclude these values from computations when determining relationships for subsequent fill steps. Therefore, it is important to perform regression data filling as early in data processing as possible so that data manipulation does not introduce derived values and bias.

The following OLS equation is used to estimate values for the dependent time series from the independent time series:

$$Y_i = \overline{Y_1} + r\frac{S_{y1}}{S_{x1}}\left[X_i - \overline{X}_1\right]$$

or

$$Y_i = a + bX_i$$

where

N_1 = concurrent or overlapping period of record (the notation N_1 is used because the MOVE2 fill technique refers to N_2, which is the number of additional points outside of N_1 in the independent time series)

$\overline{X}_1 =$ mean for independent variable for N_1 years = $\frac{\sum X_{1_i}}{N_1}$

$\overline{Y}_1 =$ mean for dependent variable for N_1 years = $\frac{\sum Y_{1_i}}{N_1}$

$S_{y1} =$ standard deviation for N_1 years = $\frac{1}{N_1 - 1}\sum\left(Y_{1i} - \overline{Y_1}\right)^2$

$S_{x1} =$ standard deviation for N_1 years = $\frac{1}{N_1 - 1}\sum\left(X_{1i} - \overline{X_1}\right)^2$

r = correlation coefficient = $\frac{N_1\sum X_{1i}Y_{1i} - \sum X_{1i}\sum Y_{1i}}{\sqrt{\left[N_1\sum X_{1_i}{}^2 - \left(\sum X_{1_i}\right)^2\right]\cdot\left[N_1\sum Y_{1i}{}^2 - \left(\sum Y_{1_i}\right)^2\right]}}$

$$b = r\frac{S_{y1}}{S_{x1}}$$

$a = \overline{Y}_1 - b\overline{X}_1$

Note that the correlation coefficient, *r*, is used to compute the slope, *b,* of the line.

A number of statistics are computed and are available for output to a table, as described below (see the `TableID` and related command parameters for how to specify the table output). Creating a statistics table and then writing the table to a file is useful for checking the analysis and software. For example, the `CompareTables()` command can be used to compare this statistics table with a verification data set that is calculated by another tool. In the following descriptions, the statistic for one equation has a name like `Mean` and monthly statistics correspondingly have a name like `Mean_1`, where `1` corresponds to January and `12` to December.

In some cases, statistics are relevant in units of the raw values, in some cases statistics are relevant in transformed (log10) units, and in some cases both are relevant. For example, if the log10 transform is used to compute the relationship, then `a` and `b` are in transformed units. However, error computations between the original data values and values that would be computed by the relationship are in the raw units (regardless of whether the data were transformed) – this allows errors to be compared between relationships using raw and translated values (the `FillMixedStation()` command uses this information to compare relationships). Consequently, the third column of the following table indicates whether statistics are provided in raw (column name uses statistic only) or transformed units (additional `_trans` added to statistic for column name), and bold indicates that where both are available only the bold version is output). If the analysis does not use a transformation, then `_trans` will be omitted from column headings. In summary, if `_trans` is shown in a column heading, then the data have been transformed and the value in the column is relevant to transformed data.

Statistics From Regression Analysis

Statistic (Table Column Name)	Involves Dependent, Independent, or Both	Statistics Output as Raw and/or Transformed Values	Description
`N1`	Both		The number (count) of non-missing data values overlapping in the dependent and independent time series.
`MeanX1`	Independent	**raw**, transformed	The mean of the independent `N1` data values.
`SX1`	Independent	**raw**, transformed	The standard deviation of the independent `N1` values.
`N2`	Independent		The number (count) of non-missing independent values outside of `N1`.
`MeanX2`	Independent	**raw**, transformed	The mean of the independent `N2` values.
`SX2`	Independent	**raw**, transformed	The standard deviation of the independent `N2` values.
`MeanY1`	Dependent	**raw**, transformed	The mean of the dependent `N1` values.
`SY1`	Dependent	**raw**, transformed	The standard deviation of the dependent `N1` values.
`NY`	Dependent		The total number of non-missing dependent values.
`MeanY`	Dependent	**raw**, transformed	The mean of the dependent `NY` values.
`SY`	Dependent	**raw**, transformed	The standard deviation of the dependent `NY` values.
`a`	Both	transformed	The intercept for the relationship equation.
`b`	Both	transformed	The slope of the relationship equation.
`R`	Both	transformed	The correlation coefficient for `N1` values.
`R2`	Both	transformed	R-squared, coefficient of determination for `N1` values.
`MeanY1est`	Dependent	**raw**, transformed	The mean for `N1` values computed from the relationship (estimate the dependent values where values were previously known).
`SY1est`	Dependent	**raw**, transformed	The standard deviation for `N1` values computed from the relationship (estimate the dependent at locations where values are known).
`RMSE`	Dependent	raw	The “room mean squared error” for `N1` overlapping values, which is a measure of the

Statistic (Table Column Name)	Involves Dependent, Independent, or Both	Statistics Output as Raw and/or Transformed Values	Description
			overall error of using the regression equation to estimate values, is calculated as: $RMSE = \sqrt{\frac{\sum \left(Y_{1_i} - Y_{1_i}'\right)^2}{N_1}}$ where Y_{1_i} is the original dependent value and Y_{1_i}' is the value estimated with the regression relationship.
`SEE`	Dependent	raw	The standard error of estimate for `N1` overlapping values, which is a measure of the overall error of using the regression equation to estimate values, calculated as: $SEE = \sqrt{\frac{\sum \left(Y_{1_i} - Y_{1_i}'\right)^2}{N_1 - 2}}$ where Y_{1_i} is the original dependent value and Y_{1_i}' is the value estimated with the regression relationship.
`SEP`	Both	raw	The standard error of prediction for `N1` overlapping values, which is a measure of the overall error of using the regression equation to estimate values,calculated as: $SEP = \sqrt{1 + \frac{1}{N_1} + \frac{(X_{1_i} - \overline{X}_1)^2}{\sum (X_{1_i} - \overline{X}_1)^2}} * SEE$ where X_{1_i} is the original independent value and $\overline{X}_1$ is the mean of the `N1` independent values.
`SESlope`	Both	transformed	The standard error (SE) of the slope (b) for `N1` overlapping values, calculated as: $SE = \frac{\sqrt{\frac{\sum (Y_{1_i} - Y'_{1_i})^2}{N_1 - 2}}}{\sqrt{\sum (X_{1_i} - \overline{X}_1)^2}}$ where X_{1_i} is the original independent value and

Comment [sam1]: Can we clarify how this is different from RMSE?

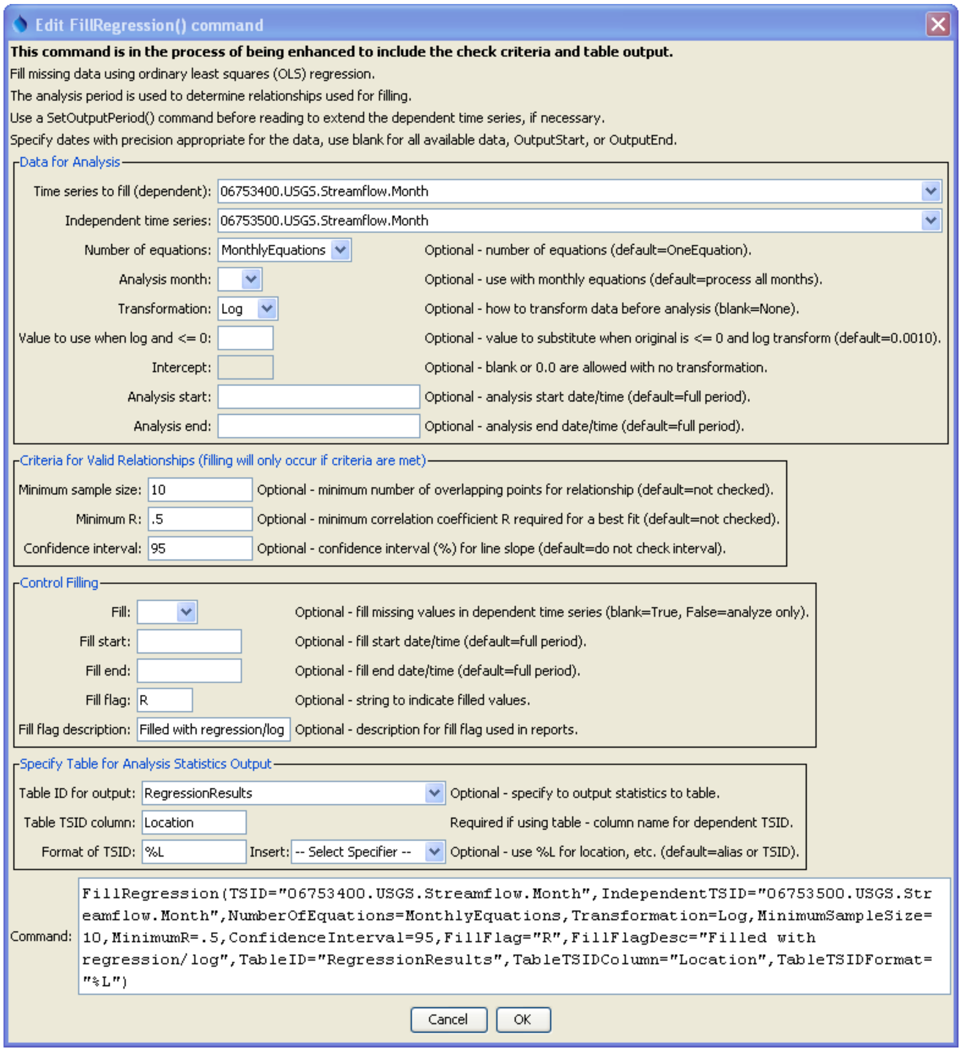
Comment [sam2]: Can we clarify how this is different from RMSE and SEE? Also, why are you (Simon) computing this for each value when this equation is an overall value. I'm concerned how individual values are used in the MSA and what if we need to apply this technique to daily data, etc.?

Statistic (Table Column Name)	Involves Dependent, Independent, or Both	Statistics Output as Raw and/or Transformed Values	Description
			$\overline{X}_1$ is the mean of the N1 independent values; Y_{1_i} is the original dependent value and Y_{1_i}' is the value estimated with the regression relationship.
TestScore	Both	transformed	b/SE
Test Quantile	Both	transformed	From the Student's T-test, which is a function of the confidence interval and degrees of freedom (DF), where DF is the degrees of freedom equal to N1 – 2 (corresponding to the intercept and the slope of the regression equation).
Test Related	Both	transformed	Will be No if TestScore < TestQuantile, indicating that the b ≠ 0 data are related, and Yes if TestScore >= TestQuantile, indicating that the data are not related. If the data are not related, then the relationship between the dependent and independent time series will not be used for filling.

Need to include a description of the T-Test and confidence interval here if possible, with a graphic. In particular, we should reference the source of the table data and describe the approach/equations (if not clear in the referenced material) so that this whole process is transparent and can be revisited.

Comment [sam3]: As discussed, it would be good to insert an explanation here and refer to other resources if appropriate via book titles/authors and/or links.

The following dialog is used to edit the command and illustrates the syntax of the command:



Edit FillRegression() command

This command is in the process of being enhanced to include the check criteria and table output.

Fill missing data using ordinary least squares (OLS) regression.

The analysis period is used to determine relationships used for filling.

Use a SetOutputPeriod() command before reading to extend the dependent time series, if necessary.

Specify dates with precision appropriate for the data, use blank for all available data, OutputStart, or OutputEnd.

Data for Analysis

Time series to fill (dependent):	06753400.USGS.Streamflow.Month	
Independent time series:	06753500.USGS.Streamflow.Month	
Number of equations:	MonthlyEquations	Optional - number of equations (default=OneEquation).
Analysis month:		Optional - use with monthly equations (default=process all months).
Transformation:	Log	Optional - how to transform data before analysis (blank=None).
Value to use when log and <= 0:		Optional - value to substitute when original is <= 0 and log transform (default=0.0010).
Intercept:		Optional - blank or 0.0 are allowed with no transformation.
Analysis start:		Optional - analysis start date/time (default=full period).
Analysis end:		Optional - analysis end date/time (default=full period).

Criteria for Valid Relationships (filling will only occur if criteria are met)

Minimum sample size:	10	Optional - minimum number of overlapping points for relationship (default=not checked).
Minimum R:	.5	Optional - minimum correlation coefficient R required for a best fit (default=not checked).
Confidence interval:	95	Optional - confidence interval (%) for line slope (default=do not check interval).

Control Filling

Fill:		Optional - fill missing values in dependent time series (blank=True, False=analyze only).
Fill start:		Optional - fill start date/time (default=full period).
Fill end:		Optional - fill end date/time (default=full period).
Fill flag:	R	Optional - string to indicate filled values.
Fill flag description:	Filled with regression/log	Optional - description for fill flag used in reports.

Specify Table for Analysis Statistics Output

Table ID for output:	RegressionResults			Optional - specify to output statistics to table.
Table TSID column:	Location			Required if using table - column name for dependent TSID.
Format of TSID:	%L	Insert:	-- Select Specifier --	Optional - use %L for location, etc. (default=alias or TSID).

Command:

```
FillRegression(TSID="06753400.USGS.Streamflow.Month",IndependentTSID="06753500.USGS.Str
eamflow.Month",NumberOfEquations=MonthlyEquations,Transformation=Log,MinimumSampleSize=
10,MinimumR=.5,ConfidenceInterval=95,FillFlag="R",FillFlagDesc="Filled with
regression/log",TableID="RegressionResults",TableTSIDColumn="Location",TableTSIDFormat=
"%L")
```

FillRegression

FillRegression() Command Editor

The command syntax is as follows:

```
FillRegression(Parameter=Value,…)
```

Command Parameters

Parameter	Description	Default
TSID	The time series identifier or alias for the time series to be filled.	None – must be specified.
IndependentTSID	The time series identifier or alias for the independent time series.	None – must be specified.
NumberOfEquations	The number of equations to use for the analysis: OneEquation or MonthlyEquations.	OneEquation
AnalysisMonth	Indicate the month to process when using monthly equations. Currently only a single month can be specified.	Process all months.
Transformation	Indicates how to transform the data before analyzing. Specify as None (previously Linear) or Log (for Log$_{10}$). If the Log option is used, zero and negative values are replaced with the value specified by the LEZeroLogValue parameter value for analysis (missing data values are ignored in the analysis).	None (no transformation).
LEZeroLogValue	Value to use for data values less than or equal to zero when using a log transformation.	.0010
Intercept	Specify as 0 to force the intercept of the best-fit line through the origin (not available for log transformation).	Parameter is optional and if specified the default is to not force the intercept through zero.
AnalysisStart	The date/time to start the analysis – use to focus on only a period appropriate from analysis. For example specify the unregulated period for streamflow.	Analyze the full period.
AnalysisEnd	The date/time to end the analysis – use to focus on only a period appropriate from analysis.	Analyze the full period.
MinimumSampleSize	The minimum number of overlapping values required to use a relationship for filling.	No limit, other than imposed by calculation of statistics.
MinimumR	The minimum correlation coefficient required to use a relationship for filling.	No check is performed.
ConfidenceInterval	A confidence interval in percent (e.g., 95) required for the slope of the relationship. The T-test is performed to ensure that the independent and dependent time series are related.	The T-test is not performed to evaluate the confidence interval.
Fill	Indicate whether fill should occur (True) or just analyze to compute statistics (False). The latter is useful for testing combinations of fill parameters prior to actually performing filling.	True
FillStart	The date/time to start filling, if other than the full time series period.	Fill the full period.
FillEnd	The date/time to end filling, if other than the full time series period.	Fill the full period.
FillFlag	A single character that will be used to flag filled data.	Filled values will not be flagged.
FillFlagDesc	Description for the fill flag, used in reports.	Automatically generated.
TableID	A table identifier for a table to receive output of the	Statistics are not written

Parameter	Description	Default
	regression analysis (statistics are described above).	to the table. Refer to the log file for information.
`TableTSIDColumn`	The name of the column in the table that contains time series identifier information. This is used to match the table with time series being analyzed so that statistics can be written to the correct row.	Required if `TableID` is specified.
`TableTSIDFormat`	The specifier used to format the time series identifier in the TableTSIDColumn. The location part of the TSID, or the time series alias is typically used.	The alias will be used if available, or otherwise the full TSID will be used.
`SEPTSID`	The time series identifier of the SEP time series, calculated for ALL values in the analysis period.	If not specified, no SEP time series will be generated.
`SEPTSAlias`	The alias to be assigned to the SEP time series.	No alias is assigned to the SEP time series.

Comment [sam4]: This is a new feature that will allow the SEP time series to be plotted and further processed. It also could be useful to create plots for this documentation to explain what is going on.

I am going to update the TableTimeSeriesMath command to allow an assignment. Then you could create a new time series, assign RMSE or other statistic to it, and use to create a graph.

The command logic is as follows, with reference to command parameters that control the process:

1. The dependent (`TSID`) and independent time series (`IndependentTSID`) are retrieved using the time series identifiers or aliases.
2. Data arrays of overlapping non-missing values are extracted from time series to be used as the samples for analysis, as specified by command parameters (analysis period specified by `AnalysisStart` and `AnalysisEnd`; transformation specified by `Transformation`, `LEZeroLogValue`, and `Intercept`; number of equations specified by `NumberOfEquations` and `AnalysisMonth`).
3. The independent and dependent statistics and relationships are calculated, computing as many of the statistics as possible (some are skipped if the sample size results in division by zero). Computing the statistics allows them to be saved in the output table for review, and is controlled by the `TableID`, `TableTSIDColumn`, and `TableTSIDFormat` parameters. If the data were transformed initially, the statistics are reported in original data units.
4. The statistics are analyzed to determine if the relationships are acceptable for filling by checking the minimum sample size (`MinimumSampleSize`), minimum correlation coefficient (`MinimumR`), and that the relationship meets the confidence interval (`ConfidenceInterval`). If monthly equations are used, then it is possible that some months can be filled but not others.
5. If `Fill=True` (the default), then the relationships that are acceptable from step 4 are used to fill the dependent time series for the period specified by the `FillStart` and `FillEnd` parameters, with `FillFlag` and `FillFlagDesc` optionally being used to indicate filled values.

Comment [sam5]: What do we need to put in the statistics table to allow checks of the Mixed Station Analysis? Do I need to include transformed values?

A sample command file to fill time series from the State of Colorado’s HydroBase is as follows:

```
# 06753400 - LONETREE CREEK AT CARR, CO.
06753400.USGS.Streamflow.Month~HydroBase
# 06753500 - LONETREE CREEK NEAR NUNN, CO.
06753500.USGS.Streamflow.Month~HydroBase
FillRegression(TSID="06753400.USGS.Streamflow.Month",
IndependentTSID="06753500.USGS.Streamflow.Month")
```